Applied Number Theory

Harald Niederreiter and Arne Winterhof

Contents

Preface

> *Et le quatrième mystère, non des moindres,*
> *est celui de la structure mathématique du monde:*
> *pourquoi et quand apparaît-elle, comment peut-on la modéliser,*
> *et comment le cerveau parvient-il à l'élaborer,*
> *à partir du chaos dans lequel nous vivons*?[1]
> E. Abécassis, *Le palimpseste d'Archimède*

The Nobel laureate Eugene Wigner coined the phrase of the "unreasonable effectiveness of mathematics" in explaining the physical world, and the novelist Eliette Abécassis expressed this amazing fact in a literary fashion and called it one of the four great mysteries of the world. Whether from the scientist's or the artist's vantage point, there is no dispute about the applicability of wide parts of mathematics. On the other hand, number theory—the purest of pure mathematics—has long resisted the temptation to become applicable, except for trivial applications such as Pythagorean triples for the construction of right angles and very simple cryptosystems. In 1940, the prominent number theorist G.H. Hardy confidently asserted in his book *A Mathematician's Apology* that he had never done anything useful and that no discovery of his is likely to make the least difference to the amenity of the world.

But things changed dramatically in the second half of the 20th century when, driven by the impetus of science and technology, entirely new areas of mathematics relying heavily on number theory were created. Nowadays number theory is implicitly present in our daily life: in the supermarket via barcodes, in our cars via GPS, in our smart phones and CD players via error-correcting codes, or when we use the Internet for online banking, to mention only a few examples.

From our perspective, there are four major areas of applications where number theory plays a fundamental role, namely cryptography, coding theory, quasi-Monte

[1] Authors' translation: And the fourth mystery, and not the least, is that of the mathematical structure of the world: why and when does it arise, how can one model it, and how does the brain manage to work it out, starting from the chaos in which we live?

Carlo methods, and pseudorandom number generation. Excellent textbooks are available for each of these areas. This book presents, for the first time, a unified account of all these applications. This allows us to delineate the manifold links and interrelations that exist between these areas. Chapters 2 to 5 cover the four main areas of applications and the last chapter reviews various additional applications of number theory, ranging from check-digit systems to quantum computation and to the organization of raster-graphics memory. We hope that this panorama of applications will inspire further research in applied number theory. In order to enhance the accessibility of the book for undergraduates, we included a brief introductory course on number theory in Chapter 1. In Chapters 2 to 5, the last section offers a glimpse of advanced results that are stated without proof and require more mathematical maturity.

We tried to minimize the prerequisites for the book. A background in number theory is not necessary, although it is certainly helpful. Elementary facts from calculus are used as a matter of course. Linear algebra appears only in a limited context, and the important special case of linear algebra over finite fields is developed from scratch. The chapters on coding theory and quasi-Monte Carlo methods are quite extensive, so that separate courses on each of these topics can be taught from them. But we believe that a single course stressing the unity of applied number theory is in better conformity with the philosophy of the book.

Writing a book is not possible without the help of many. We are particularly indebted to Professor Friedrich Pillichshammer of the University of Linz for assistance with the figures, to Professors Sheldon Axler and Ken Ribet for their comments on a preliminary version of the book, to our institutions for providing excellent research facilities, and to Edward Lear for developing limericks into an art form. The limericks at the beginning of each chapter are not credited since they were written by the first author (sorry if you find them silly). We extend our special gratitude to Ruth Allewelt and Martin Peters of Springer-Verlag for their unfailing support of our project and to our families for their patience and indulgence.

Linz, March 2015

HARALD NIEDERREITER
ARNE WINTERHOF

Chapter 1

A Review of Number Theory and Algebra

This theory of the old Greeks,
the first mathematical geeks,
is a marvel of charm and beauty,
it's number theory, this cutie,
and we are its devoted freaks.

1.1 Integer arithmetic

Elementary number theory may be regarded as a prerequisite for this book, but since we, the authors, want to be nice to you, the readers, we provide a brief review of this theory for those who already have some background on number theory and a crash course on elementary number theory for those who have not. Apart from trying to be friendly, we also follow good practice when we prepare the ground for the coming attractions by collecting some basic notation, terminology, and facts in an introductory chapter, like a playwright who presents the main characters of the play in the first few scenes. Basically, we cover only those results from elementary number theory that are actually needed in this book. For more information, there is an extensive expository literature on number theory, and if you want to read the modern classics, then we recommend the books of Hardy and Wright [62] and of Niven, Zuckerman, and Montgomery [152].

The beginning of our story is nice and easy: you count $1, 2, 3, \ldots$ *ad infinitum* and thereby you create the set $\mathbb{N}$ of all positive integers (also called natural numbers). If you throw in 0 and the negative integers $-1, -2, -3, \ldots$, then you arrive at the set $\mathbb{Z}$ of all integers with its arithmetic operations of addition, subtraction, and multiplication. You start doing number theory when you realize that there are even integers like 6 and -8 and odd integers like 9 and -5. Of course, an integer is even if and only if it is twice an integer, and from this observation it is an obvious step to the general concept of divisibility.

Definition 1.1.1 Let a and b be integers with $b \neq 0$. Then a is *divisible* by b, or equivalently b *divides* a, if there is an integer c such that $a = bc$.

There are further ways of expressing the fact that a is divisible by b, namely, a is a *multiple* of b and b is a *divisor* of a. The integers 1 and -1 are not very exciting divisors since they divide every integer. Any nonzero integer a has the trivial divisors 1, -1, a, and $-a$. Since an integer b divides $a \in \mathbb{Z}$ if and only if $-b$ divides a, one often concentrates on the positive divisors (or the *factors*) of an integer a. If $b \in \mathbb{N}$ divides $a \in \mathbb{N}$ and $b < a$, then b is called a *proper divisor* of a; if also $b > 1$, then b is called a *nontrivial divisor* (or a *nontrivial factor*) of a. The divisibility relation is transitive, in the sense that if b divides a and c divides b, then c divides a. We note again and emphasize that in divisibility relations as in Definition 1.1.1, it always goes without saying that the divisor is a nonzero integer.

Let $a \in \mathbb{Z}$ and $b \in \mathbb{N}$. Even if a is not divisible by b, we can still divide a by b, and then we get a quotient $q \in \mathbb{Z}$ and a remainder $r \in \mathbb{Z}$ with $0 \leq r < b$. Furthermore, we can write $a = qb + r$. The numbers q and r are uniquely determined. This procedure is called *division with remainder* or the *division algorithm*.

Example 1.1.2 Let us take $a = 17$ and $b = 5$. Then division with remainder yields the quotient $q = 3$ and the remainder $r = 2$, and we can write $17 = 3 \cdot 5 + 2$. If your intelligence is insulted by this example, then please ignore it.

Definition 1.1.3 For two integers a and b that are not both 0, the largest integer that divides both a and b is called the *greatest common divisor* of a and b and is denoted by $\gcd(a, b)$. Generally, for $k \geq 2$ integers $a_1, \ldots, a_k$ that are not all 0, their *greatest common divisor* $\gcd(a_1, \ldots, a_k)$ is the largest integer that divides each of $a_1, \ldots, a_k$.

It is obvious that $\gcd(a_1, \ldots, a_k)$ exists, for if without loss of generality $a_1 \neq 0$, then every divisor d of a_1 satisfies $d \leq |a_1|$.

Example 1.1.4 If $a = 12$ and $b = 18$, then the positive common divisors of a and b are 1, 2, 3, and 6, and so $\gcd(12, 18) = 6$.

Proposition 1.1.5 *If $a, b \in \mathbb{Z}$ are not both 0, then there exist $a_1, b_1 \in \mathbb{Z}$ such that*

$$\gcd(a, b) = aa_1 + bb_1.$$

Proof. Let d be the smallest element of the nonempty set

$$L = \{au + bv : u, v \in \mathbb{Z},\ au + bv > 0\}.$$

Then $d = aa_1 + bb_1 > 0$ for some $a_1, b_1 \in \mathbb{Z}$. By division with remainder, we can write $a = qd + r$ with $q, r \in \mathbb{Z}$ and $0 \leq r < d$. Then

$$r = a - qd = a - q(aa_1 + bb_1) = a(1 - qa_1) - bqb_1.$$

If we had $r > 0$, then $r \in L$, a contradiction to the definition of d. Thus $r = 0$, that is, d divides a. Similarly, one shows that d divides b.

Now let e be an arbitrary common divisor of a and b. Then e divides $aa_1+bb_1 = d$, and so $e \leq d$. Thus, d is the greatest common divisor of a and b. □

Corollary 1.1.6 *If a, b, and d are integers such that d divides ab and $\gcd(a, d) = 1$, then d divides b.*

Proof. From $\gcd(a, d) = 1$ we get $aa_1 + dd_1 = 1$ for some $a_1, d_1 \in \mathbb{Z}$ by Proposition 1.1.5. Multiplying by b, we obtain $aba_1 + dbd_1 = b$. Now d divides aba_1 and dbd_1, and so d divides b. □

Instead of looking at the common divisors of given integers, we can also consider their common multiples. If the given integers are nonzero, then they have arbitrarily large common multiples, so here the meaningful notion is the least positive common multiple.

Definition 1.1.7 For $k \geq 2$ nonzero integers $a_1, \ldots, a_k$, their *least common multiple* $\mathrm{lcm}(a_1, \ldots, a_k)$ is the smallest positive common multiple of $a_1, \ldots, a_k$.

Example 1.1.8 Let us take $a = 12$ and $b = 18$. The positive multiples of 12 are $12, 24, 36, \ldots$ and the positive multiples of 18 are $18, 36, 54, \ldots$, hence $\mathrm{lcm}(12, 18) = 36$.

You have definitely run into prime numbers like 2, 3, and 11 before, so here is the formal definition for the sake of completeness.

Definition 1.1.9 An integer $p \geq 2$ is called a *prime number* (or a *prime*) if its only positive divisors are 1 and p. If an integer $b \geq 2$ is not a prime number, then b is called a *composite number*.

Note that the integer 1 is neither a prime number nor a composite number. The prime numbers are the building blocks of the integers greater than 1, in the sense that every integer greater than 1 can be expressed in an essentially unique way as a product of prime numbers. The proof of this fundamental fact is based on the following lemma.

Lemma 1.1.10 *If a prime number p divides a product $a_1 \cdots a_s$ of integers, then p divides a_i for at least one i.*

Proof. We proceed by induction on s. The case $s = 1$ is obvious. Now suppose that p divides a product $a_1 \cdots a_{s+1}$ of $s + 1$ integers for some $s \geq 1$. If p divides a_{s+1}, then we are done. Otherwise $\gcd(a_{s+1}, p) = 1$, which implies by Corollary 1.1.6 that p divides $a_1 \cdots a_s$, and then an application of the induction hypothesis completes the proof. □

Theorem 1.1.11 (Fundamental Theorem of Arithmetic) *Every integer $b \geq 2$ can be written as a product of prime numbers and this factorization of b is unique up to the order of the prime factors.*

Proof. For the proof, we include $b = 1$ which we write as an empty product. The existence of the factorization is proved by induction on b. The case $b = 1$ is already settled. For $b \geq 2$, let d be the least divisor of b that is greater than 1. Then d is a prime number, and we apply the induction hypothesis to b/d.

In order to prove the uniqueness of the factorization into prime numbers, we again use induction on b. The case $b = 1$ is trivial. Now let $b \geq 2$ and suppose that

$$b = p_1 \cdots p_r = q_1 \cdots q_s,$$

where $p_1, \ldots, p_r, q_1, \ldots, q_s$ are prime numbers. Then p_1 divides $q_1 \cdots q_s$, and so Lemma 1.1.10 implies that p_1 divides q_i for some i with $1 \leq i \leq s$. Since q_i is a prime number, we must have $p_1 = q_i$. Thus, we can cancel p_1 against q_i and we get

$$p_2 \cdots p_r = q_1 \cdots q_{i-1} q_{i+1} \cdots q_s.$$

By the induction hypothesis, the prime factors agree on both sides up to their order, and so the prime factors of b agree up to their order. □

By collecting identical prime factors, we can write the factorization of the integer $b \geq 2$ in the form

$$b = p_1^{e_1} \cdots p_k^{e_k} = \prod_{j=1}^{k} p_j^{e_j}$$

with distinct prime numbers $p_1, \ldots, p_k$ and exponents $e_1, \ldots, e_k \in \mathbb{N}$. This is often called the *canonical factorization* of b.

Theorem 1.1.11 and the following theorem are contained in Euclid's *Elements*, the famous treatise written around 300 BC that founded geometry and number theory as rigorous mathematical disciplines.

Theorem 1.1.12 *There are infinitely many prime numbers.*

Proof. We paraphrase Euclid's original proof which is a classical gem of mathematics. Suppose there were only finitely many prime numbers and let $p_1, \dots, p_r$ be the complete list of prime numbers. Then we consider the integer $n = p_1 \cdots p_r + 1$. By Theorem 1.1.11, n has a prime factor p, and by assumption we must have $p = p_i$ for some i with $1 \le i \le r$. Then p_i divides n and $p_1 \cdots p_r$, hence p_i divides 1, which is impossible. A different proof will be presented in Remark 6.6.20. □

If we write $\prod_p$ for a product over all prime numbers, then the factorization of the integer $b \ge 2$ into prime factors can also be written in the form $b = \prod_p p^{e_p(b)}$ with uniquely determined exponents $e_p(b) \ge 0$, where only finitely many $e_p(b)$ can be positive. The case $b = 1$ can be formally included by putting $e_p(1) = 0$ for all prime numbers p. If $d \in \mathbb{N}$ is written as $d = \prod_p p^{e_p(d)}$, then d divides b if and only if $e_p(d) \le e_p(b)$ for all prime numbers p. It follows that if $a_1, \dots, a_k \in \mathbb{N}$ with $k \ge 2$, then

$$\gcd(a_1, \dots, a_k) = \prod_p p^{\min(e_p(a_1), \dots, e_p(a_k))}. \tag{1.1}$$

Similarly, we obtain

$$\operatorname{lcm}(a_1, \dots, a_k) = \prod_p p^{\max(e_p(a_1), \dots, e_p(a_k))}. \tag{1.2}$$

If $k = 2$ and the canonical factorizations of a_1 and a_2 are not readily available, then it is more efficient to compute $\gcd(a_1, a_2)$ by the Euclidean algorithm (see [152, Section 1.2] and Exercise 1.9). For $k \ge 3$ one uses the identity

$$\gcd(a_1, \dots, a_k) = \gcd(\gcd(a_1, \dots, a_{k-1}), a_k)$$

and iterations of the Euclidean algorithm.

Example 1.1.13 Let $a = 12$ and $b = 18$ as in Examples 1.1.4 and 1.1.8. Then $a = 2^2 \cdot 3^1$ and $b = 2^1 \cdot 3^2$. Thus, (1.1) shows that $\gcd(12, 18) = 2^1 \cdot 3^1 = 6$, and (1.2) shows that $\operatorname{lcm}(12, 18) = 2^2 \cdot 3^2 = 36$.

1.2 Congruences

Congruences were introduced by the "prince of mathematics" Carl Friedrich Gauss (1777–1855) in his seminal monograph *Disquisitiones Arithmeticae* written at age 24. (*A personal note*: The second author passed the Gauss memorial every day on his way to the Technical University of Braunschweig until he was 24, which may be considered the first steps towards this book. The Gauss statue holds the *Disquisitiones Arithmeticae*, the second author was reading easier lecture notes.) Congruences are an excellent tool for studying questions about divisibility and remainders.

Definition 1.2.1 Let $a, b \in \mathbb{Z}$ and let $m \in \mathbb{N}$. Then a is *congruent* to b modulo m, written $a \equiv b \pmod{m}$, if m divides the difference $a - b$. If $a - b$ is not divisible by m, then we say that a is *incongruent* to b modulo m and we write $a \not\equiv b \pmod{m}$.

The positive integer m in Definition 1.2.1 is called the *modulus* of the congruence $a \equiv b \pmod{m}$. The modulus $m = 1$ is not very exciting since $a \equiv b \pmod{1}$ for all $a, b \in \mathbb{Z}$. Therefore, interesting congruences will always involve a modulus $m \geq 2$.

Example 1.2.2 For the modulus $m = 2$, the congruence $a \equiv b \pmod{2}$ just says that the integers a and b have the same parity, that is, they are either both even or both odd.

Congruences occur in everyday life, though often in an unobtrusive form. Just take the realm of clocks and calendars as an example. If it is now 10 a.m., then five hours later it will be 3 p.m. and the reason is the congruence $10 + 5 \equiv 3 \pmod{12}$. If today is March 29, then six days later the date will be April 4 because of the congruence $29 + 6 \equiv 4 \pmod{31}$. We promise that you will see more significant applications of congruences later in the book.

To a large extent, congruences can be manipulated like equations. Given a congruence $a \equiv b \pmod{m}$, we are allowed to add or subtract the same integer on both sides and we can multiply both sides by the same integer. More generally, two congruences with the same modulus can be combined according to the following proposition.

Proposition 1.2.3 *If $a \equiv b \pmod{m}$ and $c \equiv d \pmod{m}$ with $a, b, c, d \in \mathbb{Z}$ and $m \in \mathbb{N}$, then*

$$a + c \equiv b + d \pmod{m},$$
$$a - c \equiv b - d \pmod{m},$$
$$ac \equiv bd \pmod{m}.$$

Proof. The statements about addition and subtraction of congruences are obvious from Definition 1.2.1. Finally, we note that

$$ac - bd = a(c - d) + (a - b)d,$$

and so m divides $ac - bd$ whenever m divides $a - b$ and $c - d$. □

It is a consequence of the third part of Proposition 1.2.3 that we can raise a congruence to a power. For instance, $a \equiv b \pmod{m}$ implies $a^3 \equiv b^3 \pmod{m}$. It is also easily seen that congruences are transitive, in the sense that if $a \equiv b \pmod{m}$ and $b \equiv c \pmod{m}$, then $a \equiv c \pmod{m}$. Thus, formulations like $a \equiv b \equiv c \pmod{m}$ are legitimate.

For $a \in \mathbb{Z}$ and $m \in \mathbb{N}$, division with remainder (see Section 1.1) yields $a = qm + r$ with uniquely determined $q, r \in \mathbb{Z}$ satisfying $0 \leq r < m$. It follows that $a \equiv r \pmod{m}$. The integer r is called the *least residue* of a modulo m. Every integer is thus congruent modulo m to a unique one among the integers $0, 1, \ldots, m-1$. This set of integers deserves a special name.

Definition 1.2.4 Let m be a positive integer. The set

$$Z_m := \{0, 1, \ldots, m-1\} \subset \mathbb{Z}$$

is called the *least residue system* modulo m.

There are exactly m integers in Z_m and they are pairwise incongruent modulo m. More generally, for $m \in \mathbb{N}$ we say that a set $S_m \subset \mathbb{Z}$ is a *complete residue system* modulo m if S_m contains exactly m elements and if these elements are pairwise incongruent modulo m. We again have the property that every integer is congruent modulo m to a uniquely determined element of a complete residue system modulo m.

Example 1.2.5 For an odd integer $m \geq 1$, the set $\{a \in \mathbb{Z} : -(m-1)/2 \leq a \leq (m-1)/2\}$ forms a complete residue system modulo m which is symmetric around 0. This does not work quite as well for an even integer $m \geq 2$, but in this case the set $\{a \in \mathbb{Z} : -m/2 < a \leq m/2\}$ is a complete residue system modulo m which is nearly symmetric around 0.

Proposition 1.2.6 *If $m \in \mathbb{N}$ and $a \in \mathbb{Z}$ with* $\gcd(a, m) = 1$*, then there exists a unique $c \in Z_m$ with $ac \equiv 1 \pmod{m}$.*

Proof. According to Proposition 1.1.5, there exist $a_1, m_1 \in \mathbb{Z}$ such that $aa_1 + mm_1 = 1$. It follows that $aa_1 \equiv 1 \pmod{m}$. If $c \in Z_m$ is the least residue of a_1 modulo m, then $ac \equiv aa_1 \equiv 1 \pmod{m}$.

Let $d \in Z_m$ be such that $ad \equiv 1 \pmod{m}$. Then $a(c-d) \equiv ac - ad \equiv 0 \pmod{m}$, that is, m divides $a(c-d)$. Now Corollary 1.1.6 implies that m divides $c - d$. Since $-(m-1) \leq c - d \leq m - 1$, this is possible only if $c = d$. □

Example 1.2.7 Let $m = 7$ and $a = 3$. Then $c = 5 \in Z_7$ satisfies $ac \equiv 15 \equiv 1 \pmod{7}$. In general, the integer c in Proposition 1.2.6 can be computed by means of the Euclidean algorithm (see [152, Section 1.2] and Exercise 1.12).

Definition 1.2.8 Two integers a and b that are not both 0 are said to be *coprime* (or *relatively prime*) if $\gcd(a, b) = 1$.

The following result was already known to mathematicians in ancient China and India. The formulation in the language of congruences is due to Gauss.

Theorem 1.2.9 (Chinese Remainder Theorem) *If* $m_1, \ldots, m_k \in \mathbb{N}$ *with* $k \geq 2$ *are pairwise coprime moduli and* $r_1, \ldots, r_k \in \mathbb{Z}$ *are arbitrary, then there exists a unique* $a \in Z_m$ *with* $m = m_1 \cdots m_k$ *such that*

$$a \equiv r_j \pmod{m_j} \qquad \textit{for } 1 \leq j \leq k.$$

Proof. The existence of the integer a is shown by a construction. The condition on $m_1, \ldots, m_k$ implies that $\gcd(m/m_j, m_j) = 1$ for $1 \leq j \leq k$. Hence by Proposition 1.2.6, for each j with $1 \leq j \leq k$ there exists a $c_j \in \mathbb{Z}$ with $(m/m_j)c_j \equiv 1 \pmod{m_j}$. Clearly, $(m/m_j)c_j \equiv 0 \pmod{m_i}$ for $1 \leq i, j \leq k$ with $i \neq j$. We put

$$a_0 = \sum_{j=1}^{k} (m/m_j)c_j r_j. \tag{1.3}$$

Then

$$a_0 \equiv (m/m_j)c_j r_j \equiv r_j \pmod{m_j} \qquad \text{for } 1 \leq j \leq k.$$

The same holds if we replace a_0 by $a \in Z_m$, the least residue of a_0 modulo m.

If $b \in Z_m$ with $b \neq a$ also satisfies $b \equiv r_j \pmod{m_j}$ for $1 \leq j \leq k$, then $a \equiv b \pmod{m_j}$ for $1 \leq j \leq k$. This implies that m_j divides $|a-b|$ for $1 \leq j \leq k$. As in Section 1.1, let us write $m = \prod_p p^{e_p(m)}$ and similarly for other positive integers. Since $m_1, \ldots, m_k$ are pairwise coprime, we conclude that for each prime number p we have $e_p(m) = e_p(m_j)$ for some j with $1 \leq j \leq k$. Therefore $e_p(m) = e_p(m_j) \leq e_p(|a-b|)$, and so m divides $|a-b|$. But $0 < |a-b| < m$, and thus we arrive at a contradiction. □

Example 1.2.10 The following is a version of a popular puzzle. You have a basket of eggs. When you take out three, four, or five eggs at a time, there is always one egg left, while with seven eggs at a time no egg is left. What is the least number of eggs in the basket? This word problem is equivalent to the system of congruences $a \equiv 1 \pmod{3}$, $a \equiv 1 \pmod{4}$, $a \equiv 1 \pmod{5}$, and $a \equiv 0 \pmod{7}$. The moduli 3, 4, 5, and 7 are pairwise coprime, and so we can apply the method in the proof of Theorem 1.2.9 with $m = 3 \cdot 4 \cdot 5 \cdot 7 = 420$. Note that $m/m_1 = 420/3 = 140$, and so we can take $c_1 = 2$ since $140 \cdot 2 \equiv 2 \cdot 2 \equiv 1 \pmod{3}$. Similarly, $m/m_2 = 105$ and $c_2 = 1$, and furthermore $m/m_3 = 84$ and $c_3 = 4$. We do not need c_4 since $r_4 = 0$ in (1.3). Thus, (1.3) yields $a_0 = 140 \cdot 2 + 105 \cdot 1 + 84 \cdot 4 = 721$. The least residue of 721 modulo 420 is $a = 301$, and this is the answer to the puzzle. We hope that your chickens laid more than 301 eggs because, you know, you shouldn't put *all* your eggs in one basket.

Another giant of mathematics, namely Leonhard Euler (1707–1783), introduced and employed the following number-theoretic function.

Definition 1.2.11 For $m \in \mathbb{N}$, the number of elements of Z_m that are coprime to m is denoted by $\phi(m)$. The function ϕ is called *Euler's totient function.*

Example 1.2.12 For small m, we can compute $\phi(m)$ by counting. For $m = 12$, for instance, the elements of $Z_{12} = \{0, 1, \ldots, 11\}$ that are coprime to 12 are 1, 5, 7, and 11, and so $\phi(12) = 4$. An easy general case occurs when $m = p$ is a prime number. Then all numbers $1, 2, \ldots, p-1$ in $Z_p = \{0, 1, \ldots, p-1\}$ are coprime to p, while 0 is not, and so $\phi(p) = p - 1$.

By definition, $\phi(m)$ is the number of elements of the set

$$R_m := \{a \in Z_m : \gcd(a, m) = 1\}. \tag{1.4}$$

The number $\phi(m)$ can be easily computed once the canonical factorization of $m \geq 2$ is known, as the following result shows. We note that evidently $\phi(1) = 1$.

Proposition 1.2.13 *If* $m = \prod_{j=1}^{k} p_j^{e_j}$ *is the canonical factorization of the integer* $m \geq 2$, *then*

$$\phi(m) = \prod_{j=1}^{k} \left(p_j^{e_j} - p_j^{e_j - 1}\right) = m \prod_{j=1}^{k} \left(1 - p_j^{-1}\right).$$

Proof. We first consider the case where m is a prime power, say $m = p^e$ with a prime number p and $e \in \mathbb{N}$. The elements of Z_{p^e} that are *not* coprime to p^e are exactly the multiples of p, and there are p^{e-1} of them in Z_{p^e}. Hence $\phi(p^e) = p^e - p^{e-1}$.

Now let $m = \prod_{j=1}^{k} p_j^{e_j}$ be as in the proposition and set $m_j = p_j^{e_j}$ for $1 \leq j \leq k$. We consider the map $\psi : R_m \to R_{m_1} \times \cdots \times R_{m_k}$ given by

$$\psi(a) = (\psi_1(a), \ldots, \psi_k(a)) \qquad \text{for all } a \in R_m,$$

where $\psi_j(a)$ is the least residue of a modulo m_j for $1 \leq j \leq k$. The Chinese remainder theorem (see Theorem 1.2.9) shows that ψ is bijective, and so R_m and $R_{m_1} \times \cdots \times R_{m_k}$ have the same number of elements. Therefore

$$\phi(m) = \prod_{j=1}^{k} \phi(m_j) = \prod_{j=1}^{k} \left(p_j^{e_j} - p_j^{e_j - 1}\right) = m \prod_{j=1}^{k} \left(1 - p_j^{-1}\right),$$

which is the desired result. □

Example 1.2.14 The last expression in Proposition 1.2.13 shows that in order to compute $\phi(m)$, we actually need to know only the different prime factors of m and not the full canonical factorization of m. For instance, if $m = 12$, then 2 and 3 are the different prime factors of 12, and so $\phi(12) = 12(1 - 1/2)(1 - 1/3) = 4$, which agrees with the result in Example 1.2.12.

As the first application of the number-theoretic function ϕ, we present the following classical theorem from the 18th century. Later on in Section 1.3, we will recognize this result as a special instance of a general principle in group theory.

Theorem 1.2.15 (Euler's Theorem) *If $m \in \mathbb{N}$ and $a \in \mathbb{Z}$ with $\gcd(a, m) = 1$, then*

$$a^{\phi(m)} \equiv 1 \pmod{m}.$$

Proof. We write $R_m = \{r_1, \dots, r_{\phi(m)}\}$. We multiply all elements of R_m by a to obtain the integers $ar_1, \dots, ar_{\phi(m)}$. We claim that $ar_1, \dots, ar_{\phi(m)}$ are pairwise incongruent modulo m. For if $ar_i \equiv ar_j \pmod{m}$ for some $1 \le i, j \le \phi(m)$, then multiplying the congruence by the integer c in Proposition 1.2.6, we get $r_i \equiv r_j \pmod{m}$ and so $i = j$. Moreover $ar_1, \dots, ar_{\phi(m)}$ are coprime to m because of $\gcd(a, m) = 1$, hence the least residues of $ar_1, \dots, ar_{\phi(m)}$ modulo m run through the set R_m in some order. Thus, modulo m we can compute the product of all elements of R_m in two ways to obtain

$$(ar_1) \cdots (ar_{\phi(m)}) \equiv r_1 \cdots r_{\phi(m)} \pmod{m}.$$

This means that m divides $r_1 \cdots r_{\phi(m)} \left(a^{\phi(m)} - 1\right)$. But m and $r_1 \cdots r_{\phi(m)}$ are coprime, and so m divides $a^{\phi(m)} - 1$ by Corollary 1.1.6. □

Corollary 1.2.16 (Fermat's Little Theorem) *If p is a prime number and $a \in \mathbb{Z}$ is not divisible by p, then*

$$a^{p-1} \equiv 1 \pmod{p}.$$

Proof. This follows immediately from Theorem 1.2.15 and the observation in Example 1.2.12 that $\phi(p) = p - 1$ (or the formula in Proposition 1.2.13). □

For $m \in \mathbb{N}$ and $a \in \mathbb{Z}$ with $\gcd(a, m) = 1$, we see from Theorem 1.2.15 that there is some power of a that is congruent to 1 modulo m. It is of interest to consider the smallest positive exponent for which this works.

Definition 1.2.17 For $m \in \mathbb{N}$ and $a \in \mathbb{Z}$ with $\gcd(a, m) = 1$, the least positive integer h such that $a^h \equiv 1 \pmod{m}$ is called the *multiplicative order* of a modulo m.

Example 1.2.18 Consider the prime modulus $p = 13$. Then for $a = 5$ we obtain $5^1 \equiv 5 \pmod{13}$, $5^2 \equiv 12 \pmod{13}$, $5^3 \equiv 8 \pmod{13}$, and $5^4 \equiv 1 \pmod{13}$. Thus, the multiplicative order of 5 modulo 13 is equal to 4. If we carry out the same calculation with $a = 2$, then we find that the multiplicative order of 2 modulo 13 is equal to 12, hence equal to $p - 1$. In view of Corollary 1.2.16, this is the largest possible multiplicative order that can appear modulo the prime number $p = 13$, and this situation deserves special attention.

Definition 1.2.19 Let p be a prime number and let $g \in \mathbb{Z}$ with $\gcd(g, p) = 1$. If the multiplicative order of g modulo p is equal to $p - 1$, then g is called a *primitive root* modulo p.

Remark 1.2.20 By Example 1.2.18, the integer 2 is a primitive root modulo 13. A more general principle will imply (see Corollary 1.4.33) that for every prime number p there exists a primitive root modulo p, but we will not use this result before we actually prove it.

Definition 1.2.21 Let p be an odd prime number and let a be an integer with $\gcd(a, p) = 1$. Then a is called a *quadratic residue* modulo p if there exists an integer b such that $a \equiv b^2 \pmod{p}$. If there is no such $b \in \mathbb{Z}$, then a is called a *quadratic nonresidue* modulo p.

Statements about quadratic residues can be formulated in an elegant manner by using the following notation introduced by the eminent mathematician Adrien-Marie Legendre (1752–1833).

Definition 1.2.22 Let p be an odd prime number. For all $a \in \mathbb{Z}$, the Legendre symbol $\left(\frac{a}{p}\right)$ is defined as follows. If p divides a, then $\left(\frac{a}{p}\right) = 0$. If $\gcd(a, p) = 1$, then $\left(\frac{a}{p}\right) = 1$ if a is a quadratic residue modulo p and $\left(\frac{a}{p}\right) = -1$ if a is a quadratic nonresidue modulo p.

Proposition 1.2.23 *If p is an odd prime number, then*

$$\left(\frac{a}{p}\right) \equiv a^{(p-1)/2} \pmod{p} \qquad \textit{for all } a \in \mathbb{Z}.$$

Proof. The result is trivial if p divides a, and so we can assume that $\gcd(a, p) = 1$. The argument in the proof of Theorem 1.2.15 with $m = p$ shows that for every $c \in R_p = \{1, \ldots, p-1\}$, the least residues of $c, 2c, \ldots, (p-1)c$ modulo p run through R_p in some order. Therefore there exists a unique $c' \in R_p$ with $cc' \equiv a \pmod{p}$. We pair off c with c', and then $c = c'$ occurs if and only if $\left(\frac{a}{p}\right) = 1$. Thus, if $\left(\frac{a}{p}\right) = -1$, then we can form $(p-1)/2$ distinct pairs (c, c') with $c \neq c'$ and $cc' \equiv a \pmod{p}$, and so $(p-1)! \equiv a^{(p-1)/2} \pmod{p}$. We write this congruence in the form

$$(p-1)! \equiv -\left(\frac{a}{p}\right) a^{(p-1)/2} \pmod{p}. \tag{1.5}$$

If $\left(\frac{a}{p}\right) = 1$, then $b^2 \equiv a \pmod{p}$ for some $b \in R_p$. Then $d^2 \equiv a \equiv b^2 \pmod{p}$ implies that p divides $(d-b)(d+b)$, and so $d = b$ or $d = p - b$ by Lemma 1.1.10. Now we can form $(p-3)/2$ distinct pairs (c, c') with $c \neq c'$ and $cc' \equiv a \pmod{p}$ as well as the pair $(b, p-b)$ with $b(p-b) \equiv -b^2 \equiv -a \pmod{p}$. Therefore the congruence (1.5) holds again. With $a = 1$ in (1.5) we get $(p-1)! \equiv -1 \pmod{p}$, and so

$$1 \equiv \left(\frac{a}{p}\right) a^{(p-1)/2} \pmod{p}$$

for all $a \in \mathbb{Z}$ with $\gcd(a,p) = 1$. Multiplying the last congruence by $\left(\frac{a}{p}\right)$ yields the final result. □

Proposition 1.2.24 *If p is an odd prime number, then*

$$\left(\frac{ab}{p}\right) = \left(\frac{a}{p}\right)\left(\frac{b}{p}\right) \qquad \textit{for all } a, b \in \mathbb{Z}.$$

Proof. Proposition 1.2.23 shows that

$$\left(\frac{ab}{p}\right) \equiv (ab)^{(p-1)/2} \equiv a^{(p-1)/2}b^{(p-1)/2} \equiv \left(\frac{a}{p}\right)\left(\frac{b}{p}\right) \pmod{p}.$$

Now both extreme sides of this congruence have the value 0 or ±1, and so the congruence holds if and only if equality holds (note that $p \geq 3$). □

Example 1.2.25 Let p be an odd prime number and let $a = -1$. From Proposition 1.2.23 we obtain

$$\left(\frac{-1}{p}\right) \equiv (-1)^{(p-1)/2} \pmod{p}.$$

Both sides of this congruence have the value ±1, and so we get the equality

$$\left(\frac{-1}{p}\right) = (-1)^{(p-1)/2}.$$

Thus, -1 is a quadratic residue modulo p if and only if $p \equiv 1 \pmod 4$. For instance, if $p = 13$, then $-1 \equiv 64 \equiv 8^2 \pmod{13}$.

Remark 1.2.26 For an odd prime number p, the argument after (1.5) shows that if $\left(\frac{a}{p}\right) = 1$, then there exists a unique $b \in \mathbb{Z}$ with $b^2 \equiv a \pmod p$ and $1 \leq b \leq (p-1)/2$. Hence we find all incongruent quadratic residues modulo p in the set $\{b^2 : b = 1, \ldots, (p-1)/2\}$. Therefore, there are exactly $(p-1)/2$ quadratic residues modulo p in $\{1, \ldots, p-1\}$, and consequently there are also exactly $(p-1)/2$ quadratic nonresidues modulo p in $\{1, \ldots, p-1\}$.

1.3 Groups and characters

1.3.1 Abelian groups

If you expect that in this section on groups and characters we provide a psychological study of how the character of people is affected by their social groups, then we have to disappoint you. The groups we are considering here are abelian groups in the sense of abstract algebra and the characters we are investigating are special maps between abelian groups and the set $\mathbb{C}$ of complex numbers. Moreover, we focus on abelian groups that are of number-theoretic relevance. There will be no need to consider groups that are not abelian.

If you had a course on abstract algebra, then you already know all you need to know about abelian groups for this book. For the novices in group theory, we offer a brief introduction. The study of abstract algebraic structures is best initiated with some illustrative examples.

For the theory of abelian groups, we just start with the basic set for number theory, namely the set $\mathbb{Z}$ of all integers. On $\mathbb{Z}$ we consider the binary operation of ordinary addition which assigns to every ordered pair $(a, b) \in \mathbb{Z}^2$ the sum $a + b \in \mathbb{Z}$. We have the associative law $a + (b + c) = (a + b) + c$ for all $a, b, c \in \mathbb{Z}$ and the commutative law $a + b = b + a$ for all $a, b \in \mathbb{Z}$. The integer 0 plays a special role since $a + 0 = a$ for all $a \in \mathbb{Z}$. Furthermore, for every $a \in \mathbb{Z}$ there is the integer $-a$ that can be added to it to produce 0, that is, such that $a + (-a) = 0$.

In abstract algebra we abstract (isn't that where the name of the area comes from?) from special examples, and this is what we do now. Instead of $\mathbb{Z}$ we take some set G and instead of ordinary addition in $\mathbb{Z}$ we take a binary operation $*$ on G, that is, a map that assigns to every ordered pair (a, b) of elements $a, b \in G$ an element $a * b \in G$. The properties of ordinary addition in $\mathbb{Z}$ listed in the preceding paragraph are now put forth as the axioms of an abelian group G. In particular, there must be an element of G playing the role of the integer 0, and so G will automatically be nonempty.

Definition 1.3.1 An *abelian group* is a set G together with a binary operation $*$ on G such that the following axioms hold:
(i) $a * (b * c) = (a * b) * c$ for all $a, b, c \in G$ (associative law);
(ii) $a * b = b * a$ for all $a, b \in G$ (commutative law);
(iii) there is an identity element (or neutral element) $\iota \in G$ such that $a * \iota = a$ for all $a \in G$;
(iv) for each $a \in G$, there exists an inverse element $a^{-1} \in G$ such that $a * a^{-1} = \iota$.

Remark 1.3.2 If $\iota_1, \iota_2 \in G$ are identity elements (or neutral elements) of the abelian group G, then on the one hand $\iota_1 * \iota_2 = \iota_1$ by axiom (iii) and on the other hand $\iota_1 * \iota_2 = \iota_2 * \iota_1 = \iota_2$ by axioms (ii) and (iii). It follows that $\iota_1 = \iota_2$. In other words, there is exactly one identity element (or neutral element) of G and we can speak of *the* identity element (or *the* neutral element) of G. The terminology "identity element" may suggest something like the number 1, but for instance in the case of the abelian group $\mathbb{Z}$ under ordinary addition the identity element is the integer 0. Therefore we offer also the alternative terminology "neutral element" if you feel misled by "identity element", although one has to admit that the usage of "identity element" is much more common.

Remark 1.3.3 Let $a \in G$ be given and suppose that $b, c \in G$ are inverse elements of a in the abelian group G. By using all axioms for the abelian group G, we obtain $b = b * \iota = b * (a * c) = (b * a) * c = (a * b) * c = \iota * c = c * \iota = c$. Thus, there is exactly one inverse element of a in G and we can speak of *the* inverse element of a.

The notation ι for the neutral element of an abelian group G is a bit awkward and we use it only temporarily. In practice, one often employs the additive notation $a + b$ instead of $a * b$ for the binary operation on G; but it must be emphasized that by doing so it is not assumed that the operation actually is ordinary addition of numbers. With the additive notation, it is reasonable to denote the neutral element of G by $0 \in G$. Similarly, one may write $-a \in G$ for the inverse element of $a \in G$ with the additive notation. The expression $a + (-b)$ for $a, b \in G$ is often abbreviated by $a - b$.

Possibly to confuse students, some authors prefer multiplicative notation, that is, they write ab instead of $a * b$ for the binary operation on G. In this case, it is plausible to write $1 \in G$ for the neutral element of G, or sometimes 1_G to stress the dependence on G. Again, the use of the multiplicative notation does not necessarily imply that the operation is ordinary multiplication of numbers.

Example 1.3.4 Here is an example with the multiplicative notation where the binary operation is indeed ordinary multiplication. Let $m \in \mathbb{N}$ and let U_m be the set of complex mth roots of unity. Concretely, this means that U_m consists of the complex numbers $e^{2\pi \mathrm{i} j/m}$ with $j = 0, 1, \ldots, m-1$, where $\mathrm{i} = \sqrt{-1}$ is the imaginary unit. You may want to remember here that $e^{2\pi \mathrm{i} y} = \cos(2\pi y) + \mathrm{i} \sin(2\pi y)$ for all real numbers y. The binary operation on U_m is multiplication of complex numbers. Then it is easily checked that the four axioms in Definition 1.3.1 are satisfied. The identity element of the abelian group U_m is the number 1. This is our first example of a finite abelian group, according to the following definition.

Definition 1.3.5 The abelian group G is called a *finite abelian group* if it has only finitely many elements. The number of elements of the finite abelian group G is called the *order* of G.

Example 1.3.6 Let us take a second look at the finite abelian group U_m of order m in Example 1.3.4. We put $\xi_j = e^{2\pi \mathrm{i} j/m}$ for $j \in Z_m = \{0, 1, \ldots, m-1\}$. Then Euler's identity $e^{2\pi \mathrm{i}} = 1$ yields

$$\xi_j \xi_k = e^{2\pi \mathrm{i}(j+k)/m} = e^{2\pi \mathrm{i} r/m} = \xi_r$$

for all $j, k \in Z_m$, where r is the least residue of $j + k$ modulo m. This means that the binary operation on U_m can be carried out also by adding the elements of Z_m modulo m. We thus arrive at another finite abelian group of order m, namely Z_m with the binary operation being addition modulo m. The identity element of the group Z_m is $0 \in Z_m$.

Example 1.3.7 For a fixed positive integer m, let R_m be as in (1.4). We consider the binary operation $*$ on R_m given by multiplication modulo m, that is, for $r, s \in R_m$ we let $r * s$ be the least residue of the ordinary product rs modulo m. Since $\gcd(r, m) = \gcd(s, m) = 1$ implies that $\gcd(rs, m) = 1$, we get indeed $r * s \in R_m$. It is easily checked that the first three axioms in Definition 1.3.1 are satisfied, with the identity element being the number $1 \in R_m$. The validity of the axiom (iv) follows from Proposition 1.2.6. Thus, R_m is a finite abelian group of order $\phi(m)$, where ϕ is Euler's totient function.

Now that we know a few examples of finite abelian groups, we return to the general theory of abelian groups. For elements $a_1, a_2, \ldots, a_n$ of an abelian group G with the multiplicative notation, the expression $a_1 a_2 \cdots a_n$ is unambiguous, since no matter how we insert parentheses, the expression will always represent the same element of G (thanks to the associative law). If $a_j = a$ for $1 \le j \le n$ with an element $a \in G$, then we arrive at the nth power

$$a^n = \underbrace{aa \cdots a}_{n \text{ factors}}$$

of a. It is customary to put $a^0 = 1 \in G$. With the additive notation, we get the n-fold sum

$$na = \underbrace{a + a + \cdots + a}_{n \text{ summands}}$$

of a, with the convention $0a = 0 \in G$. Usually, group theorists prefer to speak of the nth power rather than the n-fold sum.

Here is a basic result that generalizes Theorem 1.2.15. We formulate this result with the multiplicative notation.

Proposition 1.3.8 *If G is a finite abelian group of order t, then*

$$a^t = 1_G \qquad \textit{for all } a \in G.$$

Proof. We use the same idea as in the proof of Theorem 1.2.15. Let $b_1, \ldots, b_t$ be the elements of G and fix $a \in G$. Then $ab_1, \ldots, ab_t$ run again through G, for if $ab_i = ab_j$ for some $1 \le i, j \le t$, then multiplying by the inverse element a^{-1} of a we get $b_i = b_j$. It follows that

$$(ab_1) \cdots (ab_t) = b_1 \cdots b_t,$$

and so

$$a^t b_1 \cdots b_t = b_1 \cdots b_t.$$

Multiplying by the inverse element of $b_1 \cdots b_t$, we obtain $a^t = 1_G$. □

If we apply Proposition 1.3.8 to the finite abelian group R_m in Example 1.3.7, then we arrive at Theorem 1.2.15. According to Proposition 1.3.8, for a finite abelian group G there is always some power of $a \in G$ that is equal to 1_G, and so the following definition makes sense.

Definition 1.3.9 Let G be a finite abelian group and let $a \in G$. Then the least positive integer h such that $a^h = 1_G$ is called the *order* of the element a and denoted by $\mathrm{ord}(a)$.

Lemma 1.3.10 *Let G be a finite abelian group, let $a \in G$, and let $n \in \mathbb{N}$. Then $a^n = 1_G$ if and only if $\mathrm{ord}(a)$ divides n.*

Proof. With $h = \mathrm{ord}(a)$ we use division with remainder to write $n = qh + r$ with $q, r \in \mathbb{Z}$ and $0 \leq r < h$. Then

$$a^n = a^{qh+r} = (a^h)^q a^r = a^r,$$

and so $a^n = 1_G$ if and only if $a^r = 1_G$. By the definition of $\mathrm{ord}(a)$, the latter condition holds if and only if $r = 0$, that is, if and only if $\mathrm{ord}(a)$ divides n. □

Proposition 1.3.11 *If G is a finite abelian group of order t, then $\mathrm{ord}(a)$ divides t for all $a \in G$.*

Proof. This follows from Proposition 1.3.8 and Lemma 1.3.10. □

Remark 1.3.12 Let G be the finite abelian group R_m in Example 1.3.7. Then for every $a \in R_m$, the order $\mathrm{ord}(a)$ of a according to Definition 1.3.9 is the same as the multiplicative order of a modulo m (see Definition 1.2.17). It follows therefore from Proposition 1.3.11 that the multiplicative order of a modulo m always divides $\phi(m)$.

Definition 1.3.13 A finite abelian group G is called *cyclic* if there exists an element $g \in G$ such that every element of G is a power of g. The element g is called a *generator* of the finite cyclic group G. We also say that G is the cyclic group generated by g and we write $G = \langle g \rangle$.

Remark 1.3.14 If G is a finite cyclic group of order t and g is a generator of G, then $\mathrm{ord}(g) = t$. The group G consists exactly of the elements $g^0 = 1_G, g, g^2, \ldots, g^{t-1}$. A power g^n with $n \geq 0$ is a generator of G if and only if $\gcd(n, t) = 1$. It follows that G has exactly $\phi(t)$ different generators, where ϕ is Euler's totient function.

Example 1.3.15 For every $m \in \mathbb{N}$, the finite abelian group Z_m in Example 1.3.6 is cyclic since it is additively generated by the integer 1. The finite abelian group U_m in Example 1.3.4 is cyclic since it is multiplicatively generated by the complex number $e^{2\pi \mathrm{i}/m}$.

Cyclic groups often arise in the way that we take a finite abelian group G and an element $a \in G$, and we consider the finite cyclic group $\langle a \rangle$ generated by a. Since $\langle a \rangle$ is both a cyclic group and a subset of G, it is plausible to call $\langle a \rangle$ a cyclic subgroup of G. More generally, we have the following standard terminology.

Definition 1.3.16 Let G be an abelian group. A subset H of G that forms by itself a group when the binary operation on G is restricted to H is called a *subgroup* of G.

Example 1.3.17 Every abelian group G has the trivial subgroups $\{1_G\}$ and G. If U_m with $m \in \mathbb{N}$ is the finite abelian group in Example 1.3.4, then U_d is a subgroup of U_m whenever $d \in \mathbb{N}$ divides m.

Another important concept in group theory is that of a factor group. Let G be an abelian group, not necessarily finite, and let H be a subgroup of G. In the context of factor groups, we usually prefer the additive notation for the binary operation on G. For every $a \in G$, we form the *coset* (with respect to H)

$$a + H := \{a + h : h \in H\}.$$

In the multiplicative notation, we would write $aH = \{ah : h \in H\}$. For $a, b \in G$, the cosets $a + H$ and $b + H$ agree as sets if and only if $a - b \in H$. If $a + H$ and $b + H$ do not agree, then they are disjoint, for if $c \in (a + H) \cap (b + H)$, then $c - a \in H$ and $c - b \in H$, and so $a + H = c + H = b + H$.

Now we take the set of all cosets with respect to H and we introduce a binary operation on it (in the additive notation we call it the sum of cosets) as follows. For two cosets $a + H$ and $b + H$ with $a, b \in G$, their sum is defined by

$$(a + H) + (b + H) = (a + b) + H. \tag{1.6}$$

Thus, the sum of the two cosets is another coset with respect to H, as it should be. However, we need to check whether this sum is well defined, that is, if we choose arbitrary representatives $c \in a + H$ and $d \in b + H$ of the two given cosets, do we get the same sum? According to (1.6), we obtain

$$(a + H) + (b + H) = (c + H) + (d + H) = (c + d) + H.$$

But

$$(c + d) - (a + b) = \underbrace{(c - a)}_{\in H} + \underbrace{(d - b)}_{\in H} \in H$$

since H is a subgroup of G, and so indeed $(c + d) + H = (a + b) + H$. It is easy to verify that the binary operation in (1.6) satisfies all four axioms in Definition 1.3.1. The identity element is the coset $0 + H$, which is of course the subgroup H itself.

Definition 1.3.18 Let H be a subgroup of the abelian group G. Then the set of all cosets with respect to H, together with the binary operation in (1.6), forms an abelian group which is called the *factor group* G/H.

Example 1.3.19 Just for a change, let us consider an example with infinite abelian groups. The set $\mathbb{R}$ of real numbers with the binary operation of ordinary addition of real numbers is obviously an abelian group. The set $\mathbb{Z}$ of integers is a subgroup of $\mathbb{R}$. Thus, we can form the factor group $\mathbb{R}/\mathbb{Z}$. The distinct cosets with respect to $\mathbb{Z}$ are given in their canonical form by $u + \mathbb{Z}$ with the real number u running through the half-open interval $[0, 1)$. The sum of two cosets $u + \mathbb{Z}$ and $v + \mathbb{Z}$ with $u, v \in [0, 1)$ is given by $(u + v) + \mathbb{Z}$ according to (1.6). If $u + v < 1$, then the coset $(u + v) + \mathbb{Z}$ is in canonical form. If $u + v \geq 1$, then by the theory of cosets we have $(u+v)+\mathbb{Z} = (u+v-1)+\mathbb{Z}$ and the latter is in canonical form since $0 \leq u+v-1 < 1$. Thus, the cosets making up $\mathbb{R}/\mathbb{Z}$ are added by adding their representatives modulo integers. The factor group $\mathbb{R}/\mathbb{Z}$ and its multidimensional versions will play a role in the theory of quasi-Monte Carlo methods (see Subsection 4.3.2).

Example 1.3.20 Let us now start from the abelian group $\mathbb{Z}$ in Example 1.3.19 (see also the beginning of this section) and fix $m \in \mathbb{N}$. Then the set $(m) := \{km : k \in \mathbb{Z}\}$ of all multiples of m is a subgroup of $\mathbb{Z}$, and so we can form the factor group $\mathbb{Z}/(m)$. We again have a canonical form for the distinct cosets with respect to (m), namely $r + (m)$ with $r \in Z_m = \{0, 1, \ldots, m - 1\}$. Now we take a look at how the binary operation on $\mathbb{Z}/(m)$ works. The sum of two cosets $r+(m)$ and $s+(m)$ with $r, s \in Z_m$ is given by $r + s + (m)$ according to (1.6). If $r + s < m$, then the coset $r + s + (m)$ is in canonical form. If $r + s \geq m$, then $r + s + (m) = r + s - m + (m)$ and the latter is in canonical form since $0 \leq r+s-m < m$. Thus, the addition of cosets with respect to (m) is the same as addition modulo m of their representatives. We can therefore think of $\mathbb{Z}/(m)$ as another incarnation of the abelian group Z_m in Example 1.3.6. In elementary number theory, a coset with respect to (m) is also called a *residue class* modulo m.

A really fundamental application of cosets is the following beautiful result from group theory, named after the mathematician and theoretical physicist Joseph-Louis Lagrange (1736–1813).

Theorem 1.3.21 (Lagrange's Theorem) *Let G be a finite abelian group and let H be a subgroup of G. Then the order $|G|$ of G, the order $|H|$ of H, and the order $|G/H|$ of the factor group G/H are related by the identity*

$$|G| = |H| \cdot |G/H|.$$

In particular, the order of every subgroup of G divides the order of G.

Proof. We pick a coset $a_1 + H$ for some $a_1 \in G$. If $a_1 + H$ does not exhaust G, then we choose $a_2 \in G \setminus (a_1 + H)$. The cosets $a_1 + H$ and $a_2 + H$ do not agree, since $a_2 \notin a_1 + H$ and $a_2 \in a_2 + H$. Thus, by an observation about cosets above, $a_1 + H$ and $a_2 + H$ are disjoint. If the union $V = (a_1 + H) \cup (a_2 + H)$ is G, then we stop. Otherwise, we choose $a_3 \in G \setminus V$. Since G is finite, this procedure stops after a certain number s of steps, and so we arrive at cosets $a_1 + H, \dots, a_s + H$ that are pairwise disjoint and whose union is G. In other words, these cosets form a partition of G. By counting elements, we see that $|G| = |H|s$, and it is obvious that $s = |G/H|$. □

We obtain Proposition 1.3.11 as a special case of Lagrange's theorem if we choose for H the cyclic subgroup $\langle a \rangle$ of G. The following notion will lead, in Corollary 1.3.25 below, to a refinement of Proposition 1.3.8.

Definition 1.3.22 The *exponent* $E = E(G)$ of the finite abelian group G is defined by

$$E = \max_{a \in G} \operatorname{ord}(a).$$

In words, the exponent of G is the maximum order of elements of G.

Remark 1.3.23 In view of Proposition 1.3.11, the exponent E of a finite abelian group G always divides the order t of G. We have $E = t$ if and only if G is cyclic. As an example for $E < t$, consider the special case $R_8 = \{1, 3, 5, 7\}$ of the family of abelian groups R_m in Example 1.3.7. Then $\operatorname{ord}(1) = 1$ and $\operatorname{ord}(3) = \operatorname{ord}(5) = \operatorname{ord}(7) = 2$, and so $E = 2$, but obviously $t = 4$.

Proposition 1.3.24 *If G is a finite abelian group of exponent E, then* $\operatorname{ord}(a)$ *divides E for all $a \in G$.*

Proof. We consider a fixed element $a \in G$. Let p be any prime number. Then we can write $E = p^e f$ with integers $e \geq 0$ and $f \geq 1$ satisfying $\gcd(p, f) = 1$. It suffices to show that if p^r divides $\operatorname{ord}(a)$ for some integer $r \geq 0$, then we must have $r \leq e$.

We use the multiplicative notation. By Definition 1.3.22, there exists an element $b \in G$ with $\operatorname{ord}(b) = E$. Put $c = a^{\operatorname{ord}(a)/p^r}$ and $d = b^{p^e}$. Then $\operatorname{ord}(c) = p^r$ and $\operatorname{ord}(d) = f$. It follows that

$$(cd)^{p^r f} = c^{p^r f} d^{p^r f} = \left(c^{p^r}\right)^f \left(d^f\right)^{p^r} = 1_G.$$

Therefore Lemma 1.3.10 shows that $k := \operatorname{ord}(cd)$ divides $p^r f$. Next we note that $1_G = (cd)^{kf} = c^{kf} d^{kf} = c^{kf}$. Then Lemma 1.3.10 implies that p^r divides kf. Now $\gcd(p^r, f) = 1$, and so p^r divides k by Corollary 1.1.6. Similarly, we see that f divides k. Using again $\gcd(p^r, f) = 1$, we deduce that $p^r f$ divides k, and so $\operatorname{ord}(cd) = p^r f$. Finally, we invoke Definition 1.3.22 to obtain $\operatorname{ord}(cd) = p^r f \leq E = p^e f$, and so $r \leq e$ as desired. □

Corollary 1.3.25 *If G is a finite abelian group of exponent E, then*

$$a^E = 1_G \qquad \text{for all } a \in G.$$

Proof. This follows from Lemma 1.3.10 and Proposition 1.3.24. □

1.3.2 Characters

Now we know enough about group theory to talk about characters of abelian groups. An important abelian group in this context is $U = \{z \in \mathbb{C} : |z| = 1\}$, the unit circle in the complex plane, with the binary operation being ordinary multiplication of complex numbers. The abelian groups U_m in Example 1.3.4 are of course subgroups of U.

Let G be an abelian group with the multiplicative notation. Then a *character* of G is a map $\chi : G \to U$ satisfying

$$\chi(ab) = \chi(a)\chi(b) \qquad \text{for all } a, b \in G. \tag{1.7}$$

With the additive notation we require that

$$\chi(a + b) = \chi(a)\chi(b) \qquad \text{for all } a, b \in G. \tag{1.8}$$

On the right-hand sides of (1.7) and (1.8), the operation is of course ordinary multiplication of complex numbers. There are no good or bad characters of abelian groups, but there are trivial and nontrivial characters. The *trivial character* χ_0 of G is defined by $\chi_0(a) = 1$ for all $a \in G$. Any character χ of G for which $\chi(b) \neq 1$ for at least one element $b \in G$ is called a *nontrivial character* of G.

Example 1.3.26 Let G be the abelian factor group $\mathbb{R}/\mathbb{Z}$ in Example 1.3.19. For every $h \in \mathbb{Z}$, we define the map $\chi_h : \mathbb{R}/\mathbb{Z} \to U$ by

$$\chi_h(v + \mathbb{Z}) = e^{2\pi i h v} \qquad \text{for all } v \in \mathbb{R}.$$

This map is well defined, for if $u + \mathbb{Z} = v + \mathbb{Z}$ for some $u \in \mathbb{R}$, then $u - v \in \mathbb{Z}$, and so $e^{2\pi i h u} = e^{2\pi i h v}$ since $e^{2\pi i h n} = 1$ for all $n \in \mathbb{Z}$. It is obvious that (1.8) holds, and therefore χ_h is a character of $\mathbb{R}/\mathbb{Z}$. For $h = 0$ we get the trivial character χ_0 of $\mathbb{R}/\mathbb{Z}$, whereas any χ_h with $h \neq 0$ is a nontrivial character of $\mathbb{R}/\mathbb{Z}$. These characters will play a role in the theory of uniformly distributed sequences (see Subsection 4.1.1).

Example 1.3.27 For every $m \in \mathbb{N}$, let U_m be the finite abelian group in Example 1.3.4. Characters of U_m are ridiculously easy to find. Just take $\chi_1(z) = z$ for all $z \in U_m$. More generally, choose an integer h with $0 \leq h \leq m - 1$ and put $\chi_h(z) = z^h$ for all $z \in U_m$. Then (1.7) is clearly satisfied. The character χ_0 is the trivial character of U_m, and for $1 \leq h \leq m - 1$ the characters χ_h are nontrivial.

Example 1.3.28 Let p be an odd prime number and let $R_p = \{1, \ldots, p-1\}$ be the finite abelian group in Example 1.3.7 with multiplication modulo p. For $a \in R_p$, define $\eta(a) = \left(\frac{a}{p}\right)$ to be the Legendre symbol in Definition 1.2.22. Then Proposition 1.2.24 shows that η is a character of R_p. In view of Remark 1.2.26, η is a nontrivial character of R_p.

Proposition 1.3.29 *Let χ be a character of the abelian group G with identity element 1_G. Then $\chi(1_G) = 1$ and $\chi(a^{-1}) = \overline{\chi(a)}$ for every $a \in G$, where the bar denotes complex conjugation.*

Proof. With the multiplicative notation, we have $1_G 1_G = 1_G$, hence $\chi(1_G)\chi(1_G) = \chi(1_G)$ by (1.7), and so $\chi(1_G) = 1$. Furthermore, since $aa^{-1} = 1_G$ for every $a \in G$, we obtain

$$\chi(a)\chi(a^{-1}) = \chi(aa^{-1}) = \chi(1_G) = 1.$$

The complex number $\chi(a)$ has absolute value 1, and so $\chi(a^{-1}) = \overline{\chi(a)}$. □

Let us now focus on characters of finite abelian groups. The values of such characters are restricted by the following result.

Proposition 1.3.30 *Let G be a finite abelian group of exponent E. Then the values of every character of G are Eth roots of unity.*

Proof. If $a \in G$, then $a^E = 1_G$ by Corollary 1.3.25. Hence, using Proposition 1.3.29, we get $1 = \chi(1_G) = \chi(a^E) = \chi(a)^E$ for every character χ of G. □

In the case of a finite cyclic group, the characters are easy to determine, as the following example demonstrates.

Example 1.3.31 Let G be a finite cyclic group of order t and let g be a generator of G. According to Remark 1.3.14, G consists exactly of the powers g^j with $j = 0, 1, \ldots, t-1$. Proposition 1.3.30 shows that the value of a character at g is a tth root of unity, hence it is equal to $e^{2\pi \mathrm{i} h/t}$ for some integer h with $0 \leq h \leq t-1$. This value at g determines the character completely, hence we get the character χ_h of G given by

$$\chi_h(g^j) = e^{2\pi \mathrm{i} hj/t} \qquad \text{for } j = 0, 1, \ldots, t-1.$$

Note that there are exactly t different characters of G.

It is a general fact that the number of different characters of a finite abelian group G is equal to the order of G (see Theorem 1.3.36), but it requires an additional effort to establish this result. We tread carefully and we first aim to show that there are sufficiently many characters of G to separate distinct elements of G (see Lemma 1.3.33).

Lemma 1.3.32 *Let H be a subgroup of the finite abelian group G and let ψ be a character of H. Then ψ can be extended to a character of G, that is, there exists a character χ of G with $\chi(b) = \psi(b)$ for all $b \in H$.*

Proof. We use the multiplicative notation. We can suppose that $H \neq G$, for otherwise there is nothing to prove. Choose $a \in G$ with $a \notin H$. Then $H_1 = \{a^j b : j \geq 0,\ b \in H\}$ is a subgroup of G with $H \subset H_1$ since $a \in H_1$. Let m be the order of the coset aH in the factor group G/H and choose $z \in \mathbb{C}$ such that $z^m = \psi(a^m)$; note that $|z| = 1$. Now we define a map ψ_1 on H_1 by taking $b_1 \in H_1$ with $b_1 = a^j b$, $j \geq 0$, $b \in H$, and putting $\psi_1(b_1) = z^j \psi(b)$. We first have to show that ψ_1 is well defined. Thus, suppose that also $b_1 = a^k c$, $k \geq 0$, $c \in H$, where we can assume that $k > j$. Then $a^{k-j} = bc^{-1} \in H$, and so m divides $k - j$ by Lemma 1.3.10. It follows that $z^{k-j} = \psi(a^{k-j})$. Therefore

$$z^k \psi(c) = z^j z^{k-j} \psi(c) = z^j \psi(a^{k-j}) \psi(c) = z^j \psi(a^{k-j} c) = z^j \psi(b),$$

and so ψ_1 is indeed well defined.

It is obvious that ψ_1 is a character of H_1 and that $\psi_1(b) = \psi(b)$ for all $b \in H$. If $H_1 = G$, then we are done. Otherwise, we can continue the process above until, after finitely many steps, we obtain an extension of ψ to G. □

Lemma 1.3.33 *Let G be a finite abelian group and let $a_1, a_2 \in G$ with $a_1 \neq a_2$. Then there exists a character χ of G with $\chi(a_1) \neq \chi(a_2)$.*

Proof. It suffices to show that for $a = a_1 a_2^{-1} \neq 1_G$, there exists a character χ of G with $\chi(a) \neq 1$. The cyclic subgroup $H = \langle a \rangle$ of G has order $t \geq 2$. Now let ψ be the character χ_1 of H in Example 1.3.31; then $\psi(a) = e^{2\pi \mathrm{i}/t} \neq 1$. By Lemma 1.3.32, ψ can be extended to a character χ of G. □

Now we introduce a binary operation for the characters of a fixed finite abelian group G. For two characters χ and σ of G, their product $\chi\sigma$ is defined by

$$(\chi\sigma)(a) = \chi(a)\sigma(a) \qquad \text{for all } a \in G.$$

It is evident that $\chi\sigma$ is again a character of G. Let $\hat{G}$ be the set of all characters of G. Then with this product as a binary operation on $\hat{G}$, the axioms (i) and (ii) in Definition 1.3.1 are satisfied since the associative law and the commutative law hold for ordinary multiplication of complex numbers. The trivial character χ_0 of G serves as an identity element for the product of characters. Given $\chi \in \hat{G}$, its inverse element with respect to the product of characters is the character $\overline{\chi}$ of G defined by $\overline{\chi}(a) = \overline{\chi(a)}$ for all $a \in G$ (compare with Proposition 1.3.29). Altogether, $\hat{G}$ forms an abelian group under this binary operation, and since there are only finitely many choices for the values of characters of G on account of Proposition 1.3.30, $\hat{G}$ is finite. The finite abelian group $\hat{G}$ is called the *character group* (or the *dual group*) of G.

Among the results for characters, the following theorem will be most frequently used in this book.

Theorem 1.3.34 (Orthogonality Relations for Characters) *If χ is a nontrivial character of the finite abelian group G, then*

$$\sum_{a\in G}\chi(a)=0. \tag{1.9}$$

If $b \in G$ with $b \neq 1_G$, then

$$\sum_{\sigma\in\hat{G}}\sigma(b)=0. \tag{1.10}$$

Proof. We use multiplicative notation. Since χ is nontrivial, there exists an element $c \in G$ with $\chi(c) \neq 1$. Then

$$\chi(c)\sum_{a\in G}\chi(a)=\sum_{a\in G}\chi(c)\chi(a)=\sum_{a\in G}\chi(ca)=\sum_{a\in G}\chi(a),$$

because if a runs through G, then so does ca. It follows that

$$(\chi(c)-1)\sum_{a\in G}\chi(a)=0,$$

which already implies (1.9) since $\chi(c) \neq 1$.

For the second part, we introduce the function $\hat{b}$ on $\hat{G}$ by $\hat{b}(\sigma)=\sigma(b)$ for all $\sigma \in \hat{G}$. Then $\hat{b}$ is a character of $\hat{G}$. Furthermore, $\hat{b}$ is a nontrivial character since, by Lemma 1.3.33, there exists a $\chi \in \hat{G}$ with $\hat{b}(\chi)=\chi(b)\neq\chi(1_G)=1$ (recall that $b \neq 1_G$). Now we apply (1.9) to the group $\hat{G}$ and we obtain

$$\sum_{\sigma\in\hat{G}}\sigma(b)=\sum_{\sigma\in\hat{G}}\hat{b}(\sigma)=0,$$

thus proving (1.10). □

Example 1.3.35 For an odd prime number p, let η be the quadratic character of the finite abelian group R_p in Example 1.3.28. Then (1.9) yields

$$\sum_{a\in R_p}\eta(a)=0.$$

This says that the number of quadratic residues modulo p in R_p is the same as the number of quadratic nonresidues modulo p in R_p, and thus we arrive again at the result in Remark 1.2.26.

Theorem 1.3.36 *For every finite abelian group G, the number of different characters of G is equal to the order of G.*

Proof. This follows from

$$|\hat{G}| = \sum_{a \in G} \sum_{\sigma \in \hat{G}} \sigma(a) = \sum_{\sigma \in \hat{G}} \sum_{a \in G} \sigma(a) = |G|,$$

where we used (1.10) in the first identity and (1.9) in the last identity. □

1.4 Finite fields

1.4.1 Fundamental properties

This section is not a diversion into agriculture as the title may suggest, but an excursion to an area of abstract algebra called field theory which is about as important as group theory. The peculiar terminology "field" for the underlying algebraic structure is not used in all languages. For instance, in French one says *corps* and in German *Körper*, both of which mean "body". It is of course a matter of taste whether "body" captures the algebraic concept better than "field".

So, to come to the point of this section, what is a field? As for the theory of abelian groups (see Section 1.3), we start with some examples that are familiar to you. We observed in Example 1.3.19 that the set $\mathbb{R}$ of real numbers forms an abelian group under the ordinary addition of real numbers. But there is of course a second basic operation on $\mathbb{R}$, namely multiplication, and the set $\mathbb{R}^*$ of nonzero real numbers is an abelian group under this binary operation. Addition and multiplication are linked by the distributive law $u(v + w) = uv + uw$ for all $u, v, w \in \mathbb{R}$. There you already have all ingredients of a field. Analogously, the set $\mathbb{C}$ of complex numbers forms a field under the usual addition and multiplication of complex numbers. Definitely, $\mathbb{R}$ and $\mathbb{C}$ are the most popular fields in all of mathematics. Here is a third example from the hit parade of fields, namely the set $\mathbb{Q}$ of rational numbers, again of course with the ordinary addition and multiplication of rational numbers (note that the sum and the product of rational numbers are rational numbers and that the reciprocal of a nonzero rational number is again a rational number).

As in the case of abelian groups, we now take the step of abstraction. We have a set F with two binary operations which, for simplicity, we call addition and multiplication (although they are not necessarily ordinary addition and multiplication of numbers). For the result of the addition of $a, b \in F$ we write $a + b$ and for the result of the multiplication we write ab. Given the definition of an abelian group (see Definition 1.3.1), we need only three axioms to define a field.

Definition 1.4.1 A *field* is a set F together with the binary operations of addition and multiplication such that the following axioms hold:
(i) F is an abelian group under addition with identity element $0 \in F$;
(ii) $F^* := F \setminus \{0\}$ is an abelian group under multiplication with identity element $1 \in F$;
(iii) $a(b+c) = ab + ac$ for all $a, b, c \in F$ (distributive law).

It should be obvious by now that when we write 0 and 1 for an abstract field F, we do not necessarily mean the integers 0 and 1. We emphasize that every field contains at least two elements, namely the different identity elements 0 and 1. It is customary to write $-a$ for the additive inverse of $a \in F$ and a^{-1} for the multiplicative inverse of $a \in F^*$. Here are two simple properties that you use without thinking for real and complex numbers, but which hold in any field.

Lemma 1.4.2 *Let F be a field. Then:*
(i) $a0 = 0$ *for all* $a \in F$*;*
(ii) *if* $ab = 0$ *for some* $a, b \in F$*, then* $a = 0$ *or* $b = 0$*.*

Proof. (i) If $a \in F$, then $a0 = a(0+0) = a0 + a0$ by the distributive law, and so $a0 = 0$.

(ii) If $ab = 0$ and $a \neq 0$, then multiplication by a^{-1} yields $b = a^{-1}0$, and so $b = 0$ by part (i). □

Example 1.4.3 We have noted that every field contains at least the two elements 0 and 1. It is stunning, when you see this for the first time, that one can construct a field out of these two elements alone, because conventionally one thinks of fields like $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ which have infinitely many elements. Consider the set $Z_2 = \{0, 1\}$ and introduce binary operations on Z_2 by the following addition and multiplication tables.

$+$	0	1
0	0	1
1	1	0

$\cdot$	0	1
0	0	0
1	0	1

Actually, both operation tables are forced on us by the general properties of a field. Three entries in the addition table stem from the defining property of the additive identity element 0 and the fourth entry is dictated by the need for 1 to have an additive inverse. Similarly, three entries in the multiplication table stem from Lemma 1.4.2(i) and the fourth entry is due to the defining property of the multiplicative identity element 1. A second look at the addition table shows that this binary operation of addition can also be interpreted as addition modulo 2 in the set Z_2 according to Example 1.3.6. Surprisingly, this tiny set satisfies all axioms of a field (four axioms for the additive group, four axioms for the multiplicative group, plus the distributive law, so altogether nine axioms!). We already know from Example 1.3.6 that Z_2 is an abelian group under addition modulo 2. Next $Z_2^* = \{1\}$ is the trivial abelian group just consisting of the identity element 1. Finally, the distributive law $a(b+c) = ab+ac$ for all integers a, b, c implies $a(b+c) \equiv ab+ac \pmod 2$, which is the distributive law for Z_2. This is our first example of a finite field, according to the following definition.

Definition 1.4.4 A field F is called a *finite field* if it has only finitely many elements. The number of elements of a finite field F is called the *order* of F.

Every positive integer occurs as the order of some finite abelian group (see for instance Examples 1.3.4 and 1.3.6). For finite fields, there is a restriction on the possible orders: a finite field of order q exists if and only if q is a prime power. It will take some doing to prove this result. Let us start modestly by producing examples of finite fields for which the order is a prime number. To this end, we simply generalize Example 1.4.3 in an obvious manner.

Theorem 1.4.5 *For every prime number p, the least residue system modulo p given by $Z_p = \{0, 1, \ldots, p-1\}$ forms a finite field of order p under addition and multiplication modulo p.*

Proof. We know from Example 1.3.6 that Z_p is an abelian group under addition modulo p. Furthermore, $Z_p^* = \{1, \ldots, p-1\} = R_p$ is an abelian group under multiplication modulo p by Example 1.3.7. Finally, the distributive law $a(b+c) = ab+ac$ for all integers a, b, c implies $a(b+c) \equiv ab+ac \pmod p$, which is the distributive law for Z_p. □

Remark 1.4.6 A finite field for which the order is a prime number is called a *finite prime field*. For the finite prime field Z_p we use also the symbol $\mathbb{F}_p$, in line with the later notation $\mathbb{F}_q$ for a finite field of prime-power order q.

Example 1.4.7 Just for the fun of it, here are the operation tables for the finite prime field $\mathbb{F}_3 = Z_3 = \{0, 1, 2\}$.

$+$	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

$\cdot$	0	1	2
0	0	0	0
1	0	1	2
2	0	2	1

For the next step forward, we consider n-fold sums of the multiplicative identity element 1 of a field F. Concretely, for all $n \in \mathbb{N}$ we write

$$n \cdot 1 = \underbrace{1 + 1 + \cdots + 1}_{n \text{ summands}} \in F.$$

Proposition 1.4.8 *For every finite field F, there exists a least positive integer p such that $p \cdot 1 = 0 \in F$, and this integer p is a prime number.*

Proof. Consider the elements $n \cdot 1$ of F for $n \in \mathbb{N}$. Since F is finite, we must have $m \cdot 1 = n \cdot 1$ for some $m, n \in \mathbb{N}$ with $m > n$. It follows that $(m-n) \cdot 1 = m \cdot 1 - n \cdot 1 = 0 \in F$. Hence there exists a least positive integer p with $p \cdot 1 = 0 \in F$. Note that $p \geq 2$ since $1 \cdot 1 = 1 \neq 0 \in F$. Now assume that p were a composite number. Then $p = hk$ with $h, k \in \mathbb{N}$ and $1 < h, k < p$, and so $0 = p \cdot 1 = (hk) \cdot 1 = (h \cdot 1)(k \cdot 1)$. Lemma 1.4.2(ii) implies that either $h \cdot 1 = 0$ or $k \cdot 1 = 0$, but both alternatives yield contradictions to the minimality of p. □

Definition 1.4.9 The prime number p in Proposition 1.4.8 is called the *characteristic* of the finite field F. More generally, if for an arbitrary field F there exists a prime number p such that $p \cdot 1 = 0 \in F$, then p is called the *characteristic* of F.

Example 1.4.10 For every prime number p, the finite field $Z_p = \mathbb{F}_p$ in Theorem 1.4.5 has characteristic p. We remark for the sake of completeness that the fields $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ have characteristic 0 by definition, but there will be no need for us to use this terminology.

Let us consider not only $n \cdot 1$, but more generally, for every field F, for every $n \in \mathbb{N}$, and for every $a \in F$, let us put

$$n \cdot a = \underbrace{a + a + \cdots + a}_{n \text{ summands}} \in F$$

and furthermore $0 \cdot a = 0 \in F$.

The following theorem provides an important necessary condition for the order of a finite field. Later on in this section, we will prove that this condition is also sufficient. First we note a simple consequence of the definition of the characteristic.

Lemma 1.4.11 *If F is a field of characteristic p, then*

$$p \cdot a = \underbrace{a + a + \cdots + a}_{p \text{ summands}} = 0 \in F \qquad \textit{for all } a \in F.$$

Proof. If $a \in F$, then

$$p \cdot a = \underbrace{a + a + \cdots + a}_{p \text{ summands}} = a(\underbrace{1 + 1 + \cdots + 1}_{p \text{ summands}}) = a0 = 0 \in F$$

by the distributive law and Lemma 1.4.2(i). □

Theorem 1.4.12 *If F is a finite field, then the order of F is a prime power p^r, where the prime number p is the characteristic of F and r is a positive integer.*

Proof. We view F as a finite abelian group under addition. The cyclic group $\langle 1 \rangle$ generated by $1 \in F$ is a subgroup of F of order p. Therefore there exists a largest power p^r of p with $r \in \mathbb{N}$ which is the order of some subgroup H of F. Assume that $H \neq F$. Then we can choose an element $a \in F \setminus H$. A computation in the factor group F/H shows that $p(a+H) = p \cdot a + H = 0 + H$ by Lemma 1.4.11. Thus, $\operatorname{ord}(a+H)$ divides p by Lemma 1.3.10, and since p is a prime number and $a + H \neq 0 + H$, it follows that $\operatorname{ord}(a+H) = p$. By a similar argument, we obtain $\operatorname{ord}(a) = p$. Now we consider the subgroup $H_1 = \{j \cdot a + h : j \in Z_p,\ h \in H\}$ of F. If $j \cdot a + g = k \cdot a + h$ with $j, k \in Z_p$ and $g, h \in H$, then $j(a+H) = k(a+H)$, hence $j = k$ since $\operatorname{ord}(a+H) = p$, and so $g = h$. It follows that $|H_1| = |Z_p||H| = p^{r+1}$, and we get a contradiction to the definition of p^r. Thus $H = F$, and the proof is complete. □

We collect further elementary properties of a finite field F. We recall from Definition 1.4.1 that F^* denotes the multiplicative group of nonzero elements of F.

Proposition 1.4.13 *Let F be a finite field of order q. Then $a^{q-1} = 1 \in F$ for all $a \in F^*$ and $a^q = a$ for all $a \in F$.*

Proof. Since F^* is a finite abelian group of order $q-1$, the first property follows from Proposition 1.3.8. Multiplying $a^{q-1} = 1 \in F$ by $a \in F^*$, we get $a^q = a$. For $a = 0 \in F$, the identity $0^q = 0$ follows from Lemma 1.4.2(i). □

Proposition 1.4.14 *Let F be a field of characteristic p. If $a, b \in F$ and $n \in \mathbb{N}$, then*

$$(a+b)^{p^n} = a^{p^n} + b^{p^n} \quad \textit{and} \quad (a-b)^{p^n} = a^{p^n} - b^{p^n}.$$

Proof. We first take $n = 1$. In exactly the same way as for real numbers, one proves the binomial theorem

$$(a+b)^p = \sum_{j=0}^{p} \binom{p}{j} \cdot a^{p-j} b^j = a^p + \sum_{j=1}^{p-1} \binom{p}{j} \cdot a^{p-j} b^j + b^p.$$

Now

$$\binom{p}{j} = \frac{p(p-1)\cdots(p-j+1)}{1 \cdot 2 \cdot \cdots \cdot j} \equiv 0 \pmod{p}$$

for $j = 1, \ldots, p-1$ since the prime factor p in the numerator cannot be canceled. Then Lemma 1.4.11 implies that $(a+b)^p = a^p + b^p$. For arbitrary $n \in \mathbb{N}$, the first identity in the proposition is proved by induction. By what we have just shown, we obtain

$$a^{p^n} = ((a-b)+b)^{p^n} = (a-b)^{p^n} + b^{p^n},$$

and the second identity follows. □

1.4.2 Polynomials

You are of course familiar with polynomials over the real numbers and the complex numbers. The general theory of polynomials proceeds in complete analogy. For an arbitrary field F, a *polynomial* (over F) in the variable (or indeterminate) x is a formal expression

$$f(x) = \sum_{j=0}^{n} a_j x^j = a_n x^n + \cdots + a_1 x + a_0 \tag{1.11}$$

with an integer $n \geq 0$ and coefficients $a_j \in F$ for $0 \leq j \leq n$. The set of all polynomials over F in the variable x is denoted by $F[x]$. If $a_j = 0 \in F$ for $0 \leq j \leq n$ in (1.11), then we have the zero polynomial $0 \in F[x]$. If $a_n \neq 0 \in F$ in (1.11), then a_n is called the *leading coefficient* of $f(x)$. If $a_n = 1 \in F$, then the polynomial $f(x)$ is called *monic*. The coefficient a_0 in (1.11) is the *constant term* of $f(x)$. If we write a nonzero polynomial $f(x) \in F[x]$ as in (1.11) with leading coefficient $a_n \neq 0 \in F$, then the *degree* of $f(x)$ is defined by $\deg(f(x)) = n$. Various conventions are in use for the degree of the zero polynomial $0 \in F[x]$. For the time being, we put $\deg(0) = -\infty$. Later on in the book, we will utilize other conventions for $\deg(0)$.

Polynomials over F are added and multiplied just like polynomials over $\mathbb{R}$, the only difference being that the arithmetic of coefficients is the arithmetic in F. If we then inspect the axioms for a field (see Definitions 1.4.1 and 1.3.1), then we realize that only one of the axioms fails, namely the existence of a multiplicative inverse. For instance, the multiplicative inverse of $x \in \mathbb{R}[x]$ would be $\frac{1}{x}$, but this is a rational function and not a polynomial.

An algebraic structure that satisfies all axioms for a field except the existence of a multiplicative inverse is called a commutative ring with identity, or simply a *ring*. Thus, we speak of the *polynomial ring* $F[x]$. Another famous example of a ring that is not a field is the ring $\mathbb{Z}$ of integers, with the binary operations being ordinary addition and multiplication of integers. Here, for instance, the multiplicative inverse of $2 \in \mathbb{Z}$ would be $\frac{1}{2}$, but this is not an integer. We get a field if we pass from $\mathbb{Z}$ to $\mathbb{Q}$.

The product of polynomials behaves very nicely with respect to the degree. For nonzero polynomials $f(x), g(x) \in F[x]$ with leading coefficients a_n of x^n and b_m of x^m, respectively, the coefficient of x^{n+m} in $f(x)g(x)$ is $a_n b_m \neq 0 \in F$ and this is the leading coefficient of $f(x)g(x)$. Therefore

$$\deg(f(x)g(x)) = \deg(f(x)) + \deg(g(x)). \tag{1.12}$$

If at least one of $f(x)$ and $g(x)$ is $0 \in F[x]$, then this formula holds as well, with the obvious interpretation $n + (-\infty) = -\infty$ for all $n \in \mathbb{N} \cup \{0, -\infty\}$. It follows from (1.12) that if $f(x) \neq 0 \in F[x]$ and $g(x) \neq 0 \in F[x]$, then also $f(x)g(x) \neq 0 \in F[x]$. In other words, the polynomial ring $F[x]$ satisfies the property in Lemma

1.4.2(ii). A ring with this additional property is called an *integral domain*. Clearly, the ring $\mathbb{Z}$ of integers is also an integral domain.

There is no good analog of (1.12) for the sum of polynomials. The best we can say is that

$$\deg(f(x)+g(x)) \leq \max(\deg(f(x)), \deg(g(x)))$$

for all $f(x), g(x) \in F[x]$, with the obvious interpretation if some of the degrees are $-\infty$.

Example 1.4.15 Let F be the finite prime field $\mathbb{F}_3$ and let $f(x) = x^3+1 \in \mathbb{F}_3[x]$ and $g(x) = 2x^3+x+1 \in \mathbb{F}_3[x]$. Over $\mathbb{R}$ we would have $f(x)g(x) = 2x^6+x^4+3x^3+x+1$, but since now the coefficients have to be computed modulo 3, we obtain

$$f(x)g(x) = 2x^6 + x^4 + x + 1 \in \mathbb{F}_3[x].$$

We see that (1.12) is of course satisfied. Similarly, over $\mathbb{R}$ we would have $f(x) + g(x) = 3x^3 + x + 2$, but over $\mathbb{F}_3$ we get

$$f(x) + g(x) = x + 2 \in \mathbb{F}_3[x].$$

Here we have a case where $\deg(f(x)+g(x)) = 1$ is smaller than $\max(\deg(f(x)), \deg(g(x))) = 3$.

There is a theory of divisibility for polynomials over F similar to that for integers (see Section 1.1). We say that a nonzero polynomial $g(x) \in F[x]$ *divides* a polynomial $f(x) \in F[x]$ if there exists a polynomial $h(x) \in F[x]$ such that $f(x) = g(x)h(x)$. We have the same alternative ways of expressing divisibility as in Definition 1.1.1 and the paragraph following it; for instance, we speak of the *divisor* $g(x)$ of a polynomial $f(x)$ if $g(x)$ divides $f(x)$. Moreover, $g(x)$ is a *proper divisor* of $f(x)$ if $g(x)$ divides $f(x)$ and $\deg(g(x)) < \deg(f(x))$. We see in this last definition that the analog of the condition $b < a$ for integers is the condition $\deg(g(x)) < \deg(f(x))$ for polynomials. From this it is clear how the *division algorithm* (or *division with remainder*) works for polynomials: for any nonzero $g(x) \in F[x]$ and any $f(x) \in F[x]$ there exist uniquely determined polynomials $l(x), r(x) \in F[x]$ such that $f(x) = l(x)g(x) + r(x)$ and $\deg(r(x)) < \deg(g(x))$.

We have to be a bit careful when we define the greatest common divisor of polynomials. It is not enough to say that "greatest" means "largest degree" in the context of polynomials, for if $g(x)$ divides $f(x)$ in $F[x]$ and $c \in F^*$, then $cg(x)$ also divides $f(x)$.

Definition 1.4.16 Let F be a field. For $k \geq 2$ polynomials $f_1(x), \ldots, f_k(x) \in F[x]$ that are not all 0, their *greatest common divisor* $\gcd(f_1(x), \ldots, f_k(x))$ is the uniquely determined monic polynomial over F of largest degree that divides each of $f_1(x), \ldots, f_k(x)$. If $k = 2$ and $\gcd(f_1(x), f_2(x)) = 1$, then we say that $f_1(x)$ and $f_2(x)$ are *coprime* (or *relatively prime*).

We can be brief with the greatest common divisor of polynomials because the statements and proofs are completely analogous to those for the greatest common divisor of integers in Section 1.1.

Proposition 1.4.17 (i) *For all $f(x), g(x) \in F[x]$ that are not both* 0*, there exist $f_1(x), g_1(x) \in F[x]$ such that*

$$\gcd(f(x), g(x)) = f(x)f_1(x) + g(x)g_1(x).$$

(ii) *Let $f(x), g(x), h(x) \in F[x]$. If $h(x)$ divides $f(x)g(x)$ and* $\gcd(f(x), h(x)) = 1$*, then $h(x)$ divides $g(x)$.*

Proof. (i) Instead of the set L in the proof of Proposition 1.1.5, we consider

$$M = \{f(x)l(x) + g(x)m(x) : \ l(x), m(x) \in F[x]\}.$$

By the hypothesis, M contains a nonzero polynomial over F, and so we can choose a monic polynomial $d(x) \in M$ of least degree. In fact, $d(x)$ is uniquely determined, for if there were a monic $d_1(x) \in M$ with $\deg(d_1(x)) = \deg(d(x))$ and $d_1(x) \neq d(x)$, then $d(x) - d_1(x) \in M$ and $0 \leq \deg(d(x) - d_1(x)) < \deg(d(x))$, hence multiplying $d(x) - d_1(x)$ by the multiplicative inverse of its leading coefficient we get a contradiction to the choice of $d(x)$. As in the proof of Proposition 1.1.5, one shows that $d(x)$ divides $f(x)$ and $g(x)$. Similarly, any common divisor of $f(x)$ and $g(x)$ divides $d(x)$, and so $d(x) = \gcd(f(x), g(x)) = f(x)f_1(x) + g(x)g_1(x)$ for some $f_1(x), g_1(x) \in F[x]$. This argument shows also that $\gcd(f(x), g(x))$ is uniquely determined. In an analogous way, it can be seen that the greatest common divisor of $k \geq 2$ polynomials over F, not all 0, is uniquely determined.

(ii) Proceed as in the proof of Corollary 1.1.6. □

Definition 1.4.18 Let F be a field. For $k \geq 2$ nonzero polynomials $f_1(x), \ldots, f_k(x) \in F[x]$, their *least common multiple* $\operatorname{lcm}(f_1(x), \ldots, f_k(x))$ is the uniquely determined monic polynomial over F of least degree that is a common multiple of $f_1(x), \ldots, f_k(x)$.

The role of the prime numbers in the ring $\mathbb{Z}$ of integers is played by the irreducible polynomials in the ring $F[x]$.

Definition 1.4.19 Let F be a field. A polynomial $p(x) \in F[x]$ with $\deg(p(x)) \geq 1$ is said to be *irreducible over F* (or *irreducible in $F[x]$*) if it allows no factorization $p(x) = f(x)g(x)$ with $f(x), g(x) \in F[x]$, $1 \leq \deg(f(x)) < \deg(p(x))$, and $1 \leq \deg(g(x)) < \deg(p(x))$. A polynomial in $F[x]$ of positive degree that is not irreducible over F is called *reducible over F* (or *reducible in $F[x]$*).

Remark 1.4.20 It is important to emphasize irreducible (or reducible) *over F* since the irreducibility or reducibility of a given polynomial depends heavily on the field under consideration. For instance, the polynomial $x^2 - 2 \in \mathbb{Q}[x]$ is irreducible over the field $\mathbb{Q}$ of rational numbers, but $x^2 - 2 = (x + \sqrt{2})(x - \sqrt{2})$ is reducible over the field $\mathbb{R}$ of real numbers.

Example 1.4.21 A linear polynomial, that is, a polynomial of degree 1, over any field F is always irreducible over F. Now consider the polynomial $p(x) = x^2 + x + 1$ over the finite prime field $\mathbb{F}_2$. The only possibility for a nontrivial factorization of $p(x)$ over $\mathbb{F}_2$ is $p(x) = f(x)g(x)$ with $f(x), g(x) \in \mathbb{F}_2[x]$ and $\deg(f(x)) = \deg(g(x)) = 1$. Since there are only two linear polynomials over $\mathbb{F}_2$, namely x and $x + 1$, we can simply try all choices for $f(x)$ and $g(x)$. We have $x \cdot x = x^2$, $x(x + 1) = x^2 + x$, and $(x + 1)^2 = x^2 + 1$, and consequently $p(x) = x^2 + x + 1$ is irreducible over $\mathbb{F}_2$. Extensive tables of monic irreducible polynomials over the finite prime fields $\mathbb{F}_2$, $\mathbb{F}_3$, $\mathbb{F}_5$, and $\mathbb{F}_7$ can be found in the books [102, Chapter 10] and [103, Chapter 10].

Theorem 1.4.22 *For every field F, each polynomial $f(x) \in F[x]$ with $\deg(f(x)) \geq 1$ has a canonical factorization*

$$f(x) = c \prod_{j=1}^{k} p_j(x)^{e_j},$$

where $c \in F^*$, $e_1, \dots, e_k \in \mathbb{N}$, and $p_1(x), \dots, p_k(x)$ are distinct monic irreducible polynomials in $F[x]$. This factorization is unique up to the order of the factors.

Proof. Proceed as in the proof of Theorem 1.1.11, using in particular Proposition 1.4.17(ii) in the proof of uniqueness. □

Example 1.4.23 In $\mathbb{F}_2[x]$ we have $x^6 + 1 = (x^3 + 1)^2$ and $x^3 + 1 = (x + 1)(x^2 + x + 1)$. Then $x^6 + 1 = (x + 1)^2(x^2 + x + 1)^2$ is the canonical factorization of $x^6 + 1$ in $\mathbb{F}_2[x]$ since $x + 1$ and $x^2 + x + 1$ are irreducible over $\mathbb{F}_2$ (see Example 1.4.21).

Analogs of the formulas (1.1) and (1.2) can be established on the basis of the canonical factorization for polynomials over a field F. If we write $\prod_{p(x)}$ for a product over all monic irreducible polynomials over F, then the factorization of a nonzero polynomial $f(x) \in F[x]$ into monic irreducible factors over F can be written in the form

$$f(x) = c \prod_{p(x)} p(x)^{e_{p(x)}(f(x))}$$

with $c \in F^*$ and exponents $e_{p(x)}(f(x)) \geq 0$, where only finitely many $e_{p(x)}(f(x))$ can be positive. If $f_1(x), \dots, f_k(x) \in F[x]$ are $k \geq 2$ nonzero polynomials, then

$$\gcd(f_1(x), \dots, f_k(x)) = \prod_{p(x)} p(x)^{\min(e_{p(x)}(f_1(x)), \dots, e_{p(x)}(f_k(x)))}$$

and

$$\operatorname{lcm}(f_1(x), \ldots, f_k(x)) = \prod_{p(x)} p(x)^{\max(e_{p(x)}(f_1(x)), \ldots, e_{p(x)}(f_k(x)))}.$$

For two fields F and K for which $F \subseteq K$ and the addition and multiplication in F are the addition and multiplication in K restricted to F, we say that F is a *subfield* of K or that K is an *extension field* of F. For instance, $\mathbb{Q}$ is a subfield of $\mathbb{R}$ and $\mathbb{C}$ is an extension field of $\mathbb{R}$. Because of Lagrange's theorem (see Theorem 1.3.21), a finite prime field (see Remark 1.4.6) cannot contain any strictly smaller subfield. For $f(x) \in F[x]$ and an element α in some extension field K of F, the function value $f(\alpha) \in K$ is obtained by the substitution $x = \alpha$.

Now the discussion becomes even more agricultural since we will talk about roots in fields and stoop down to count the roots in a field.

Definition 1.4.24 Let F be a field. Then an element α in an extension field of F is called a *root* (or a *zero*) of $f(x) \in F[x]$ if $f(\alpha) = 0 \in F$.

Lemma 1.4.25 *Let F be a field and let K be an extension field of F. Then $\alpha \in K$ is a root of $f(x) \in F[x]$ if and only if the linear polynomial $x - \alpha$ divides $f(x)$ in $K[x]$.*

Proof. By the division algorithm in $K[x]$, we can write $f(x) = l(x)(x - \alpha) + \beta$ with $l(x) \in K[x]$ and $\beta \in K$. Therefore $f(\alpha) = \beta$, and the desired result follows from this identity. □

If $\alpha \in K$ and $f(x) \in F[x]$ are as in Lemma 1.4.25 and α is a root of $f(x)$, then it can happen that not only $x - \alpha$, but also a higher power of $x - \alpha$ divides $f(x)$ in $K[x]$. For a nonzero polynomial $f(x) \in F[x]$, there is a largest power $(x - \alpha)^m$ that divides $f(x)$ in $K[x]$, and then m is called the *multiplicity* of the root α. If $m = 1$, then α is called a *simple root* (or a *simple zero*) of $f(x)$, and if $m \geq 2$, then α is called a *multiple root* (or a *multiple zero*) of $f(x)$.

Example 1.4.26 Let $f(x) = x^6 + 1 \in \mathbb{F}_2[x]$. Then Example 1.4.23 shows that $(x + 1)^2$ divides $f(x)$ in $\mathbb{F}_2[x]$, but $(x + 1)^3$ does not. Therefore $\alpha = 1 \in \mathbb{F}_2$ is a multiple root of $f(x)$ with multiplicity 2.

Theorem 1.4.27 *Let F be a field and let $f(x) \in F[x]$ with* $\deg(f(x)) = n \geq 0$. *Then in every extension field of F, the polynomial $f(x)$ has at most n roots, counting multiplicities.*

Proof. The case $n = 0$ is trivial since then $f(x)$ consists only of a nonzero constant term. Now let $n \geq 1$ and let the distinct elements $\alpha_1, \ldots, \alpha_r \in K$ be roots of $f(x)$ in an extension field K of F, with respective multiplicities $m_1, \ldots, m_r$. Then

$(x-\alpha_1)^{m_1}, \ldots, (x-\alpha_r)^{m_r}$ occur as factors in the canonical factorization of $f(x)$ in $K[x]$, and so $\prod_{j=1}^{r}(x-\alpha_j)^{m_j}$ divides $f(x)$ in $K[x]$. By comparing degrees, we get $\sum_{j=1}^{r} m_j \leq n$. □

A characterization of multiple roots of polynomials is obtained by borrowing the concept of derivative from calculus. For $f(x) \in F[x]$ given by (1.11), its *derivative* $f'(x)$ is defined in the expected way as

$$f'(x) = \sum_{j=1}^{n} (j \cdot a_j) x^{j-1} \in F[x],$$

where $j \cdot a_j$ is the j-fold sum of a_j for $1 \leq j \leq n$. The usual rules for derivatives, such as the product rule, hold for every field F.

Proposition 1.4.28 *Let F be a field and let K be an extension field of F. Then $\alpha \in K$ is a multiple root of the nonzero polynomial $f(x) \in F[x]$ if and only if α is a root of both $f(x)$ and $f'(x)$.*

Proof. If $\alpha \in K$ is a root of $f(x)$, then $f(x) = (x-\alpha)g(x)$ for some $g(x) \in K[x]$ by Lemma 1.4.25. The product rule yields $f'(x) = g(x) + (x-\alpha)g'(x)$, and so $f'(\alpha) = g(\alpha)$. By definition, α is a multiple root of $f(x)$ if and only if $g(\alpha) = 0$, and so the desired result follows. □

Corollary 1.4.29 *Let F be a field. Then a nonzero polynomial $f(x) \in F[x]$ with* $\gcd(f(x), f'(x)) = 1$ *has only simple roots in every extension field of F.*

Proof. This is an immediate consequence of Proposition 1.4.28. □

1.4.3 Constructions of finite fields

Now we are ready to construct general finite fields, following in the footsteps of Evariste Galois (1811–1832). Galois is the romantic hero of mathematics: a brilliant mathematician who revolutionized algebra, who passionately engaged in French politics, and who died at age 21 in what seemed to be a duel. If you want to read a really good book on a mathematical genius, then we recommend the biographical novel on Galois by Petsinis [159]. An excellent source book on the work of Galois is [123]. The major achievements of Galois were Galois theory in algebra and the theory of finite fields (also called "Galois fields" in his honor) in number theory and algebra.

So far, the only finite fields we know are finite prime fields (see Theorem 1.4.5), and this was also the state of affairs before Galois came along. An arbitrary finite field F has a prime number p as its characteristic (see Proposition 1.4.8 and Definition 1.4.9) and it must contain the n-fold sums $n \cdot 1$ for all $n \in \mathbb{N}$. Therefore F has a copy of $\mathbb{F}_p = Z_p$ as a subfield. Galois posits a universe in which to operate, and in the modern interpretation this would be the algebraic closure $\overline{\mathbb{F}_p}$ of $\mathbb{F}_p$, that is, the field $\overline{\mathbb{F}_p}$ consisting of all roots of all polynomials over $\mathbb{F}_p$ of positive degree. This is analogous to the step from $\mathbb{R}$ to $\mathbb{C}$ which is taken in order to accommodate all roots of all polynomials over $\mathbb{R}$ of positive degree. We refer to [173, Chapter 2] for more information on the algebraic closure $\overline{F}$ of an arbitrary field F. In particular, we use the fact that each polynomial over F of positive degree always has a root in $\overline{F}$.

We know from Theorem 1.4.12 that the order of a finite field is necessarily a prime power. The following result of Galois provides the crucial converse.

Theorem 1.4.30 *For every prime power q, there exists a finite field of order q.*

Proof. Let $q = p^r$ with a prime number p and $r \in \mathbb{N}$. Consider the polynomial $f(x) = x^q - x \in \mathbb{F}_p[x]$ and let F be the set of all roots of $f(x)$ in $\overline{\mathbb{F}_p}$. We have $f'(x) = (q \cdot 1)x^{q-1} - 1 = -1$ since $q \cdot 1 = 0 \in \mathbb{F}_p$, and therefore $\gcd(f(x), f'(x)) = \gcd(x^q - x, -1) = 1$. It follows then from Corollary 1.4.29 that $f(x)$ has only simple roots in $\overline{\mathbb{F}_p}$, and so F has exactly $\deg(f(x)) = q$ elements.

It remains to verify that F is a subfield of $\overline{\mathbb{F}_p}$. We proceed by Definition 1.4.1 and we show first that F is a subgroup of the additive group $\overline{\mathbb{F}_p}$. For this it suffices to prove that if $\alpha, \beta \in F$, then also $\alpha - \beta \in F$. Indeed, Proposition 1.4.14 yields

$$(\alpha - \beta)^q = (\alpha - \beta)^{p^r} = \alpha^{p^r} - \beta^{p^r} = \alpha^q - \beta^q = \alpha - \beta,$$

and so $\alpha - \beta \in F$. Finally, we show that the set F^* of nonzero elements of F is a subgroup of the multiplicative group $\overline{\mathbb{F}_p}^*$. Again, it suffices to prove that if $\alpha, \beta \in F^*$, then $\alpha\beta^{-1} \in F^*$. Now $(\alpha\beta^{-1})^q = \alpha^q(\beta^{-1})^q = \alpha^q(\beta^q)^{-1} = \alpha\beta^{-1}$, and the proof is complete. □

In a paper published posthumously, Gauss criticized the approach by Galois and disparaged "the liberty that some younger mathematicians have taken by introducing imaginary quantities". Therefore alternative approaches to the construction of finite fields were developed, and we will present one such approach later in Remark 1.4.44.

Somehow the question of how to best construct finite fields is moot since it is a theorem (see [103, Theorem 2.5]) that all finite fields of the same order are basically identical, in the sense that they have the same algebraic structure and just differ by the names or symbols that we assign to their elements. Therefore we can speak of *the* finite field $\mathbb{F}_q$ of order q. For the manifold applications of finite fields that we will encounter in the present book, it is in principle immaterial which description of $\mathbb{F}_q$ is used. All that matters is that there exists a set $\mathbb{F}_q$ of size q which forms a field. Next we note a simple relationship between finite fields.

Proposition 1.4.31 *Let q be a prime power and let $n \in \mathbb{N}$. Then $\mathbb{F}_q$ is a subfield of $\mathbb{F}_{q^n}$, or in other words, $\mathbb{F}_{q^n}$ is an extension field of $\mathbb{F}_q$.*

Proof. In view of the proof of Theorem 1.4.30, it suffices to show that the polynomial $x^q - x \in \mathbb{F}_p[x]$ divides the polynomial $x^{q^n} - x \in \mathbb{F}_p[x]$, where p is the characteristic of $\mathbb{F}_q$. But this is readily seen: the integer $q-1$ divides the integer $q^n - 1$, hence $x^{q-1} - 1 \in \mathbb{F}_p[x]$ divides $x^{q^n-1} - 1 \in \mathbb{F}_p[x]$, and so $x^q - x \in \mathbb{F}_p[x]$ divides $x^{q^n} - x \in \mathbb{F}_p[x]$. □

Here is a remarkable property of finite fields which is useful in many applications of finite fields. We remind you that we defined cyclic groups in Definition 1.3.13.

Theorem 1.4.32 *For every finite field $\mathbb{F}_q$, the multiplicative group $\mathbb{F}_q^*$ of nonzero elements of $\mathbb{F}_q$ is cyclic.*

Proof. Let $E = E(\mathbb{F}_q^*)$ be the exponent of the finite abelian group $\mathbb{F}_q^*$ (see Definition 1.3.22). Then Corollary 1.3.25 shows that $a^E = 1 \in \mathbb{F}_q$ for all $a \in \mathbb{F}_q^*$, that is, every $a \in \mathbb{F}_q^*$ is a root of the polynomial $x^E - 1 \in \mathbb{F}_q[x]$. Theorem 1.4.27 implies that $q - 1 \leq E$. On the other hand, it is trivial that $E \leq q-1$. Hence $E = q-1$, which means that there exists an element $g \in \mathbb{F}_q^*$ with $\operatorname{ord}(g) = q-1$. □

Corollary 1.4.33 *For every prime number p, there exists a primitive root modulo p.*

Proof. The group $\mathbb{F}_p^*$ is cyclic by Theorem 1.4.32, and (the least residue modulo p of) a primitive root modulo p is nothing else but a generator of the cyclic group $\mathbb{F}_p^*$. □

Definition 1.4.34 For a finite field $\mathbb{F}_q$, every generator of the cyclic group $\mathbb{F}_q^*$ is called a *primitive element* of $\mathbb{F}_q$.

Remark 1.4.35 It follows from Remark 1.3.14 that, for every prime power q, there are exactly $\phi(q-1)$ primitive elements of $\mathbb{F}_q$, where ϕ is Euler's totient function. In particular, there are exactly $\phi(p-1)$ primitive roots modulo p in the least residue system modulo p.

Now we offer some relaxation with a brief interlude about general fields. For a given field F, let α be an element of an extension field of F such that α is a root of a polynomial over F of positive degree; we call such an element *algebraic over F*. We are interested in the set of all polynomials over F that have α as a root. This set contains a polynomial that is singled out by the following result.

Proposition 1.4.36 *Let F be a field and let α be an element of an extension field of F such that α is algebraic over F. Then there exists a uniquely determined monic polynomial $m(x) \in F[x]$ of least degree having α as a root.*

Proof. Since α is algebraic over F, there exists a polynomial $f(x) \in F[x]$ with $\deg(f(x)) \geq 1$ and $f(\alpha) = 0$. Among all such polynomials $f(x)$, we can choose one called $m(x) \in F[x]$ of least degree and we can make $m(x)$ monic. It remains to show that if $g(x) \in F[x]$ is monic with $g(\alpha) = 0$ and $\deg(g(x)) = \deg(m(x))$, then $g(x) = m(x)$. But if we had $g(x) \neq m(x)$, then we get an easy contradiction to the construction of $m(x)$ by considering the polynomial $c^{-1}(g(x) - m(x)) \in F[x]$, where c is the leading coefficient of $g(x) - m(x)$. □

Definition 1.4.37 Let F be a field and let α be an element of an extension field of F such that α is algebraic over F. Then the uniquely determined polynomial $m(x) \in F[x]$ in Proposition 1.4.36 is called the *minimal polynomial* of α over F.

Proposition 1.4.38 *Let F be a field, let α be an element of an extension field of F such that α is algebraic over F, and let $m(x) \in F[x]$ be the minimal polynomial of α over F. Then a polynomial $f(x) \in F[x]$ satisfies $f(\alpha) = 0$ if and only if $m(x)$ divides $f(x)$ in $F[x]$.*

Proof. By the division algorithm, we can write $f(x) = l(x)m(x) + r(x)$ with $l(x), r(x) \in F[x]$ and $\deg(r(x)) < \deg(m(x))$. This implies that $f(\alpha) = 0$ if and only if $r(\alpha) = 0$. The definition of $m(x)$ shows that $r(\alpha) = 0$ if and only if $r(x)$ is the zero polynomial, that is, if and only if $m(x)$ divides $f(x)$ in $F[x]$. □

Proposition 1.4.39 *Let F be a field and let α be an element of an extension field of F such that α is algebraic over F. Then the minimal polynomial of α over F is irreducible over F.*

Proof. This is easy enough: if $m(x) \in F[x]$ is the minimal polynomial of α over F and if we had $m(x) = f(x)g(x)$ with $f(x), g(x) \in F[x]$, $1 \leq \deg(f(x)) < \deg(m(x))$, and $1 \leq \deg(g(x)) < \deg(m(x))$, then $0 = m(\alpha) = f(\alpha)g(\alpha)$, and so $f(\alpha) = 0$ or $g(\alpha) = 0$ by Lemma 1.4.2(ii); but this is in any case a contradiction to the definition of $m(x)$. □

Remark 1.4.40 Let $F = \mathbb{F}_q$ and consider the extension field $\mathbb{F}_{q^n}$ with $n \in \mathbb{N}$. Then every $\alpha \in \mathbb{F}_{q^n}$ satisfies $\alpha^{q^n} = \alpha$ by Proposition 1.4.13, or in other words, α is a root of the polynomial $x^{q^n} - x \in \mathbb{F}_q[x]$. Therefore α is algebraic over $\mathbb{F}_q$. Consequently, the results in Propositions 1.4.38 and 1.4.39 apply to all elements of all finite fields.

Given again an arbitrary field F, let the elements $\alpha_1, \ldots, \alpha_k$ from some extension field of F be algebraic over F. Then $F(\alpha_1, \ldots, \alpha_k)$ is by definition the smallest field containing $F, \alpha_1, \ldots, \alpha_k$; more precisely, we take an extension field K of F with

$\alpha_1, \ldots, \alpha_k \in K$ and then $F(\alpha_1, \ldots, \alpha_k)$ is the intersection of all subfields of K that contain $F, \alpha_1, \ldots, \alpha_k$. If $k = 1$, that is, if we form $F(\alpha)$ with an algebraic element $\alpha \in K$ over F, then $F(\alpha)$ is called a *simple extension field* of F.

Example 1.4.41 Let $F = \mathbb{F}_q$ and $\mathbb{F}_{q^n}$ with $n \in \mathbb{N}$ be as in Remark 1.4.40. As an algebraic element $\alpha \in \mathbb{F}_{q^n}$ over $\mathbb{F}_q$, we choose a primitive element α of $\mathbb{F}_{q^n}$ which exists by Theorem 1.4.32 and Definition 1.4.34. Then the field $\mathbb{F}_q(\alpha)$ is a subfield of $\mathbb{F}_{q^n}$ containing $\mathbb{F}_q$ and all powers of α. But the powers of α exhaust $\mathbb{F}_{q^n}^*$, and so $\mathbb{F}_q(\alpha) = \mathbb{F}_{q^n}$. Therefore in the world of finite fields, every finite extension field of every finite field is a simple extension field.

The situation described in Example 1.4.41 arises quite frequently in applications, and so the following special terminology is used in this case.

Definition 1.4.42 Let $\mathbb{F}_q$ be a finite field. A polynomial over $\mathbb{F}_q$ is called a *primitive polynomial* over $\mathbb{F}_q$ if it is the minimal polynomial over $\mathbb{F}_q$ of some primitive element of some finite extension field of $\mathbb{F}_q$.

We note that since a primitive polynomial over $\mathbb{F}_q$ is a minimal polynomial over $\mathbb{F}_q$, a primitive polynomial over $\mathbb{F}_q$ is automatically monic and irreducible over $\mathbb{F}_q$. It is comforting to know that there exist primitive polynomials of any positive degree.

Proposition 1.4.43 *For every finite field $\mathbb{F}_q$ and for every $n \in \mathbb{N}$, there exists a primitive polynomial over $\mathbb{F}_q$, and so in particular a monic irreducible polynomial over $\mathbb{F}_q$, of degree n.*

Proof. Choose a primitive element α of the extension field $\mathbb{F}_{q^n}$ and let $p(x) \in \mathbb{F}_q[x]$ be the minimal polynomial of α over $\mathbb{F}_q$. Put $d = \deg(p(x))$ and consider the subset K of $\mathbb{F}_{q^n}$ consisting of all elements $\sum_{j=0}^{d-1} c_j \alpha^j$ with $c_0, c_1, \ldots, c_{d-1} \in \mathbb{F}_q$. It is obvious that K is a subgroup of the group $\mathbb{F}_{q^n}$ under addition. Actually, we want to prove that K is a subfield of $\mathbb{F}_{q^n}$. First we show that if $\beta, \gamma \in K$, then also $\beta\gamma \in K$. We can write $\beta = f(\alpha)$ and $\gamma = g(\alpha)$ with $f(x), g(x) \in \mathbb{F}_q[x]$, $\deg(f(x)) < d$, and $\deg(g(x)) < d$. By the division algorithm, $f(x)g(x) = l(x)p(x) + r(x)$ with $l(x), r(x) \in \mathbb{F}_q[x]$ and $\deg(r(x)) < d$. Then $\beta\gamma = f(\alpha)g(\alpha) = r(\alpha)$ since $p(\alpha) = 0$, and so $\beta\gamma \in K$. Next we prove that if $\beta \in K$ with $\beta \neq 0$, then $\beta^{-1} \in K$. We write $\beta = f(\alpha)$ with a nonzero polynomial $f(x) \in \mathbb{F}_q[x]$ satisfying $\deg(f(x)) < d$. Then $\gcd(f(x), p(x)) = 1$ since $p(x)$ is irreducible over $\mathbb{F}_q$ by Proposition 1.4.39. Thus, Proposition 1.4.17(i) shows that we can write $1 = f(x)f_1(x) + p(x)p_1(x)$ with $f_1(x), p_1(x) \in \mathbb{F}_q[x]$, where we can achieve $\deg(f_1(x)) < d$ by subtracting a suitable multiple of $p(x)$. Substituting $x = \alpha$, we get $1 = \beta f_1(\alpha)$, and so $\beta^{-1} = f_1(\alpha) \in K$. Hence K is indeed a subfield of $\mathbb{F}_{q^n}$.

Note that K contains $\mathbb{F}_q$ and α, and therefore $\mathbb{F}_q(\alpha) \subseteq K \subseteq \mathbb{F}_{q^n}$. Now $\mathbb{F}_q(\alpha) = \mathbb{F}_{q^n}$ by Example 1.4.41, and so $K = \mathbb{F}_{q^n}$. Next we observe that for every $\beta \in K$, the representation $\beta = \sum_{j=0}^{d-1} c_j \alpha^j$ with $c_0, c_1, \dots, c_{d-1} \in \mathbb{F}_q$ is unique, for if β also had a different representation of this type, then we would get an immediate contradiction to the definition of $p(x)$ as the minimal polynomial of α over $\mathbb{F}_q$. Thus, K has exactly q^d elements, hence $q^d = q^n$, and so $d = n$ and $\deg(p(x)) = d = n$. □

We present a table listing for each $n = 1, \dots, 15$ a primitive polynomial $p(x)$ over $\mathbb{F}_2$ of degree n. This table is extracted from [103, Chapter 10, Table D]. Many more examples of primitive polynomials can be found in the tables in [103, Chapter 10].

n	$p(x)$
1	$x + 1$
2	$x^2 + x + 1$
3	$x^3 + x + 1$
4	$x^4 + x + 1$
5	$x^5 + x^2 + 1$
6	$x^6 + x + 1$
7	$x^7 + x + 1$
8	$x^8 + x^4 + x^3 + x^2 + 1$
9	$x^9 + x^4 + 1$
10	$x^{10} + x^3 + 1$
11	$x^{11} + x^2 + 1$
12	$x^{12} + x^6 + x^4 + x + 1$
13	$x^{13} + x^4 + x^3 + x + 1$
14	$x^{14} + x^5 + x^3 + x + 1$
15	$x^{15} + x + 1$

Remark 1.4.44 Here is the long-awaited alternative construction of general finite fields. Basically, we find all ingredients of this construction in the proof of Proposition 1.4.43. There is also an analogy with the factor group $\mathbb{Z}/(m)$ in Example 1.3.20. Let $q = p^r$ with a prime number p and $r \in \mathbb{N}$. According to Proposition 1.4.43, we can choose an irreducible polynomial $f(x) \in \mathbb{F}_p[x]$ over $\mathbb{F}_p$ of degree r. Note that $\mathbb{F}_p[x]$ is an abelian group with the binary operation being addition of polynomials and that the set $(f(x)) := \{l(x)f(x) : l(x) \in \mathbb{F}_p[x]\}$ of all multiples of $f(x)$ is a subgroup of $\mathbb{F}_p[x]$. Therefore we can form the factor group $\mathbb{F}_p[x]/(f(x))$. The distinct elements of this factor group are the cosets $g(x) + (f(x))$, where $g(x) \in \mathbb{F}_p[x]$ and $\deg(g(x)) < r$. Hence $\mathbb{F}_p[x]/(f(x))$ is a finite abelian group of order p^r. Now we make a field out of $\mathbb{F}_p[x]/(f(x))$ by introducing a multiplication for cosets in the obvious manner: we define

$$(g(x) + (f(x)))(h(x) + (f(x))) = g(x)h(x) + (f(x)) \tag{1.13}$$

for all $g(x), h(x) \in \mathbb{F}_p[x]$. Using the trick in the proof of Proposition 1.2.3, we see that this multiplication is well defined. The same argument as in the proof of Proposition 1.4.43, now with $f(x)$ in the role of $p(x)$, shows that every nonzero element of $\mathbb{F}_p[x]/(f(x))$ has an inverse element with respect to the binary operation in (1.13). Therefore $\mathbb{F}_p[x]/(f(x))$ forms a finite field of order $q = p^r$. Gauss would have been satisfied with this construction of finite fields, as it involves no "imaginary quantities". By the way, the existence of an irreducible polynomial $f(x) \in \mathbb{F}_p[x]$ over $\mathbb{F}_p$ of degree r can be shown also by a combinatorial method (see [103, Section 3.2]), without recourse to the arguments in the proof of Proposition 1.4.43.

Example 1.4.45 Let us construct the finite field $\mathbb{F}_4$ according to the procedure in Remark 1.4.44. As the irreducible polynomial $f(x)$ over $\mathbb{F}_2$ we take $f(x) = x^2+x+1 \in \mathbb{F}_2[x]$ (see Example 1.4.21). We have exactly four cosets in $\mathbb{F}_2[x]/(f(x))$, namely $0 + (f(x))$, $1 + (f(x))$, $x + (f(x))$, and $x + 1 + (f(x))$, which we abbreviate by $\overline{0}$, $\overline{1}$, $\overline{x}$, and $\overline{x+1}$, respectively. By recalling how the arithmetic operations with cosets work, we obtain the following addition and multiplication tables.

$+$	$\overline{0}$	$\overline{1}$	$\overline{x}$	$\overline{x+1}$
$\overline{0}$	$\overline{0}$	$\overline{1}$	$\overline{x}$	$\overline{x+1}$
$\overline{1}$	$\overline{1}$	$\overline{0}$	$\overline{x+1}$	$\overline{x}$
$\overline{x}$	$\overline{x}$	$\overline{x+1}$	$\overline{0}$	$\overline{1}$
$\overline{x+1}$	$\overline{x+1}$	$\overline{x}$	$\overline{1}$	$\overline{0}$

$\cdot$	$\overline{0}$	$\overline{1}$	$\overline{x}$	$\overline{x+1}$
$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$
$\overline{1}$	$\overline{0}$	$\overline{1}$	$\overline{x}$	$\overline{x+1}$
$\overline{x}$	$\overline{0}$	$\overline{x}$	$\overline{x+1}$	$\overline{1}$
$\overline{x+1}$	$\overline{0}$	$\overline{x+1}$	$\overline{1}$	$\overline{x}$

For instance, the entry in the lower right corner of the multiplication table is obtained by noting that $(x+1)\cdot(x+1) = x^2+1$ and $x^2+1+(f(x)) = x+(f(x))$ since $f(x)$ divides $x^2+1-x = x^2+x+1$ in $\mathbb{F}_2[x]$. The set $\{\overline{0}, \overline{1}\}$ represents the subfield $\mathbb{F}_2$ of $\mathbb{F}_4$.

Remark 1.4.46 The approach in Remark 1.4.44 can be used also for the construction of finite extension fields of an arbitrary finite field $\mathbb{F}_q$. Given a positive integer k, we can choose an irreducible polynomial $f(x) \in \mathbb{F}_q[x]$ over $\mathbb{F}_q$ of degree k (see Proposition 1.4.43). With $(f(x))$ being the set of all multiples of $f(x)$ in $\mathbb{F}_q[x]$, we form the factor group $\mathbb{F}_q[x]/(f(x))$. This is a finite abelian group of order q^k, and by introducing a multiplication of cosets as in (1.13), we obtain a finite field of order q^k. We can identify each coset with respect to $(f(x))$ with a uniquely determined polynomial $r(x) \in \mathbb{F}_q[x]$ satisfying $\deg(r(x)) < k$, and the operations for these polynomials are carried out modulo $f(x)$. We speak of the *residue class field* $\mathbb{F}_q[x]/(f(x))$.

A theory of congruences (see Section 1.2) for polynomials over $\mathbb{F}_q$ can be developed with every nonzero modulus $m(x) \in \mathbb{F}_q[x]$. We say that $g_1(x) \in \mathbb{F}_q[x]$ is *congruent* to $g_2(x) \in \mathbb{F}_q[x]$ modulo $m(x)$, and we write $g_1(x) \equiv g_2(x) \pmod{m(x)}$, provided that $m(x)$ divides the difference $g_1(x) - g_2(x)$ in $\mathbb{F}_q[x]$; otherwise we say that $g_1(x)$ is *incongruent* to $g_2(x)$ modulo $m(x)$. Obviously, the basic properties of congruences in Proposition 1.2.3 hold here as well. A coset in $\mathbb{F}_q[x]$ with respect to $(m(x)) := \{l(x)m(x) : l(x) \in \mathbb{F}_q[x]\}$ is also called a *residue class* modulo $m(x)$. This explains the terminology "residue class field" in Remark 1.4.46. By defining addition and multiplication of residue classes modulo $m(x)$ in the obvious manner (compare with Remarks 1.4.44 and 1.4.46), we obtain the *residue class ring* $\mathbb{F}_q[x]/(m(x))$.

If $g(x) \equiv r(x) \pmod{m(x)}$ with $\deg(r(x)) < \deg(m(x))$, then $r(x) \in \mathbb{F}_q[x]$ is called the *least residue* of $g(x) \in \mathbb{F}_q[x]$ modulo $m(x) \in \mathbb{F}_q[x]$. A set of $q^{\deg(m(x))}$ polynomials over $\mathbb{F}_q$ that are pairwise incongruent modulo $m(x)$ is called a *complete residue system* modulo $m(x)$. An easy example of a complete residue system modulo $m(x)$ is given by the *least residue system* modulo $m(x)$, that is, the set of all $r(x) \in \mathbb{F}_q[x]$ with $\deg(r(x)) < \deg(m(x))$.

It is a remarkable phenomenon that once we know a root of a polynomial over a finite field $\mathbb{F}_q$, then further roots of that polynomial can be generated in a very simple manner. Concretely, let $f(x) = \sum_{j=0}^{n} c_j x^j \in \mathbb{F}_q[x]$ with $c_j \in \mathbb{F}_q$ for $0 \le j \le n$ and $\deg(f(x)) = n \ge 1$, and let α be a root of $f(x)$ in some extension field of $\mathbb{F}_q$. Then using Propositions 1.4.14 and 1.4.13, we get

$$0 = f(\alpha)^q = \Big(\sum_{j=0}^{n} c_j \alpha^j\Big)^q = \sum_{j=0}^{n} c_j^q \alpha^{jq} = \sum_{j=0}^{n} c_j \alpha^{jq} = f(\alpha^q),$$

and so α^q is also a root of $f(x)$. If we feed α^q into this formula, then we obtain that $(\alpha^q)^q = \alpha^{q^2}$ is a root of $f(x)$. In the end, all elements α^{q^s} with $s = 0, 1, \dots$ are roots of $f(x)$. Obviously, these elements cannot all be distinct since $f(x)$ has at most n roots by Theorem 1.4.27. A particularly nice situation occurs in the case where $f(x)$ is irreducible over $\mathbb{F}_q$, because then the different ones among the elements α^{q^s}, $s = 0, 1, \dots$, yield exactly all roots of $f(x)$.

Proposition 1.4.47 *Let $\mathbb{F}_q$ be a finite field and let $f(x) \in \mathbb{F}_q[x]$ be irreducible over $\mathbb{F}_q$ with $\deg(f(x)) = k$. Then $f(x)$ has a root α in the finite extension field $\mathbb{F}_{q^k}$, all roots of $f(x)$ are simple, and the roots of $f(x)$ are exactly the k distinct elements $\alpha, \alpha^q, \alpha^{q^2}, \dots, \alpha^{q^{k-1}}$ of $\mathbb{F}_{q^k}$.*

Proof. As a model for the finite field $\mathbb{F}_{q^k}$ we take the residue class field $\mathbb{F}_q[x]/(f(x))$ in Remark 1.4.46. Let α denote the coset $x + H$ with respect to $H := (f(x))$. Then by the way the arithmetic operations in $\mathbb{F}_q[x]/(f(x))$ are defined, we obtain

$$f(\alpha) = f(x + H) = f(x) + H = 0 + H,$$

and so $\alpha \in \mathbb{F}_{q^k}$ is a root of $f(x)$. As we have seen, it follows that $\alpha, \alpha^q, \alpha^{q^2}, \dots, \alpha^{q^{k-1}}$ are roots of $f(x)$. Since $f(x)$ can have at most k roots by Theorem 1.4.27, it suffices now to show that $\alpha, \alpha^q, \alpha^{q^2}, \dots, \alpha^{q^{k-1}}$ are distinct. So suppose we had $\alpha^{q^i} = \alpha^{q^j}$ for some $i, j \in \mathbb{Z}$ with $0 \le i < j \le k-1$. By raising this identity to the power q^{k-j}, we get $\alpha^{q^d} = \alpha^{q^k} = \alpha$ with $d = k + i - j$, where we used $\alpha \in \mathbb{F}_{q^k}$ and Proposition 1.4.13 in the second step. We infer from the proof of Theorem 1.4.30 that $\alpha \in \mathbb{F}_{q^d}$, and so $\mathbb{F}_q(\alpha) \subseteq \mathbb{F}_{q^d}$. On the other hand, the definition of α as the coset $x + H$ shows that $\mathbb{F}_q(\alpha) = \mathbb{F}_q[x]/(f(x)) = \mathbb{F}_{q^k}$, and so $\mathbb{F}_{q^k} \subseteq \mathbb{F}_{q^d}$. This is a contradiction to $d = k + i - j < k$. □

1.4.4 Trace map and characters

We introduce an important map from a finite field to a subfield which will turn out to be useful, for instance, in the construction of characters of finite fields later in this subsection. In order to simplify the notation, we write $F = \mathbb{F}_q$ for a given finite field and $K = \mathbb{F}_{q^n}$ with $n \in \mathbb{N}$ for a finite extension field of F. We start from an element $\alpha \in K$ and we consider the element $\gamma = \sum_{j=0}^{n-1} \alpha^{q^j}$ which, as it stands, lies in K. But now we observe that by Propositions 1.4.14 and 1.4.13 we obtain

$$\gamma^q = \big(\sum_{j=0}^{n-1} \alpha^{q^j}\big)^q = \sum_{j=0}^{n-1} \alpha^{q^{j+1}} = \sum_{j=1}^{n-1} \alpha^{q^j} + \alpha^{q^n} = \sum_{j=1}^{n-1} \alpha^{q^j} + \alpha = \gamma,$$

and so the argument in the proof of Theorem 1.4.30 shows that $\gamma \in \mathbb{F}_q = F$.

Definition 1.4.48 Let $F = \mathbb{F}_q$ be an arbitrary finite field and let $K = \mathbb{F}_{q^n}$ with $n \in \mathbb{N}$ be a finite extension field of F. Then the *trace* $\mathrm{Tr}_{K/F}(\alpha)$ of $\alpha \in K$ over F is defined by

$$\mathrm{Tr}_{K/F}(\alpha) = \sum_{j=0}^{n-1} \alpha^{q^j} \in F.$$

Example 1.4.49 Let $q = 2$ and $n = 2$ in Definition 1.4.48, so that $F = \mathbb{F}_2$ and $K = \mathbb{F}_4$. Then by the addition and multiplication tables in Example 1.4.45, we obtain

$$\mathrm{Tr}_{K/F}(\overline{x}) = \overline{x} + \overline{x}^2 = \overline{x} + \overline{x+1} = \overline{1} \in F.$$

Theorem 1.4.50 *Let $F = \mathbb{F}_q$ and $K = \mathbb{F}_{q^n}$ with $n \in \mathbb{N}$. Then the trace map $\mathrm{Tr}_{K/F} : K \to F$ has the following properties:*
(i) $\mathrm{Tr}_{K/F}(\alpha + \beta) = \mathrm{Tr}_{K/F}(\alpha) + \mathrm{Tr}_{K/F}(\beta)$ *for all $\alpha, \beta \in K$;*
(ii) $\mathrm{Tr}_{K/F}(c\alpha) = c\,\mathrm{Tr}_{K/F}(\alpha)$ *for all $c \in F$ and $\alpha \in K$;*
(iii) *for every $c \in F$, there are exactly q^{n-1} elements $\alpha \in K$ with $\mathrm{Tr}_{K/F}(\alpha) = c$, and so in particular the map $\mathrm{Tr}_{K/F}$ is surjective.*

Proof. (i) This is an immediate consequence of Proposition 1.4.14.

(ii) This follows from $c^{q^j} = c$ for all $c \in F$ and all integers $j \geq 0$, which is in turn deduced from Proposition 1.4.13.

(iii) For every $c \in F$, let $N(c)$ be the number of $\alpha \in K$ with $\mathrm{Tr}_{K/F}(\alpha) = c$. Note that $\mathrm{Tr}_{K/F}(\alpha) = \sum_{j=0}^{n-1} \alpha^{q^j} = c$ if and only if α is a root of the polynomial $\sum_{j=0}^{n-1} x^{q^j} - c \in F[x]$ of degree q^{n-1}. Hence Theorem 1.4.27 shows that $N(c) \leq q^{n-1}$ for all $c \in F$. Consequently, we obtain

$$q^n = \sum_{c \in F} N(c) \leq \sum_{c \in F} q^{n-1} = q \cdot q^{n-1} = q^n.$$

It follows that we must have equality throughout, and so $N(c) = q^{n-1}$ for all $c \in F$. □

The last task we set ourselves in this chapter is to determine the characters of a given finite field $\mathbb{F}_q$. According to the definition of a field in Definition 1.4.1, there are actually two abelian groups that are relevant in this context, namely the additive group $\mathbb{F}_q$ (that is, $\mathbb{F}_q$ with the binary operation being addition) and the multiplicative group $\mathbb{F}_q^*$ (that is, the set $\mathbb{F}_q^*$ of nonzero elements of $\mathbb{F}_q$ with the binary operation being multiplication). Both abelian groups $\mathbb{F}_q$ and $\mathbb{F}_q^*$ are of course finite, and so the general theory of characters of finite abelian groups in Subsection 1.3.2 applies.

Let us first consider the additive group $\mathbb{F}_q$. A character of this group $G = \mathbb{F}_q$ is a map $\chi : G \to U = \{z \in \mathbb{C} : |z| = 1\}$ satisfying (1.8). The basic tool for the construction of such a character is the trace map $\mathrm{Tr}_{\mathbb{F}_q/\mathbb{F}_p} : \mathbb{F}_q \to \mathbb{F}_p$, where p is the characteristic of $\mathbb{F}_q$ and $\mathbb{F}_p$ is the finite prime field contained in $\mathbb{F}_q$. In order to avoid awkward notation, we abbreviate this trace map by Tr in the following discussion. As usual, we identify $\mathbb{F}_p$ with $Z_p = \{0, 1, \ldots, p-1\} \subset \mathbb{Z}$ under the arithmetic modulo p. Now we choose an element $c \in \mathbb{F}_q$ and we put

$$\chi_c(a) = e^{2\pi \mathrm{i}\mathrm{Tr}(ca)/p} \qquad \text{for all } a \in \mathbb{F}_q. \tag{1.14}$$

Then it follows from Theorem 1.4.50(i) that the map $\chi_c : \mathbb{F}_q \to U$ is a character of the additive group $\mathbb{F}_q$. Instead of "character of the additive group $\mathbb{F}_q$", we shall henceforth use the terminology *additive character* of $\mathbb{F}_q$.

Theorem 1.4.51 *The additive characters of the finite field $\mathbb{F}_q$ are exactly given by the maps χ_c in* (1.14) *with c running through $\mathbb{F}_q$.*

Proof. We have already seen that each map χ_c with $c \in \mathbb{F}_q$ is an additive character of $\mathbb{F}_q$. Furthermore, we know from Theorem 1.3.36 that there are exactly q different additive characters of $\mathbb{F}_q$. Therefore it suffices to prove that the maps χ_b and χ_c are different whenever $b, c \in \mathbb{F}_q$ with $b \neq c$. By Theorem 1.4.50(iii) there exists an element $d \in \mathbb{F}_q$ with $\mathrm{Tr}(d) = 1 \in \mathbb{F}_p$. With $a = (b-c)^{-1}d \in \mathbb{F}_q$ we then deduce from (1.14) that

$$\frac{\chi_b(a)}{\chi_c(a)} = e^{2\pi \mathrm{i}\mathrm{Tr}((b-c)a)/p} = e^{2\pi \mathrm{i}\mathrm{Tr}(d)/p} = e^{2\pi \mathrm{i}/p} \neq 1,$$

and so $\chi_b(a) \neq \chi_c(a)$. □

Now we turn to the multiplicative group $\mathbb{F}_q^*$, a character of which is called a *multiplicative character* of $\mathbb{F}_q$. This is actually the easier case since we can just collect the fruits of earlier labor. The point is that $\mathbb{F}_q^*$ is a finite cyclic group of order $q-1$ by Theorem 1.4.32 and that the characters of all finite cyclic groups were already determined in Example 1.3.31.

Theorem 1.4.52 *Let g be a fixed primitive element of the finite field $\mathbb{F}_q$. Then for each integer $h = 0, 1, \dots, q-2$, the map $\psi_h : \mathbb{F}_q^* \to U$ given by*

$$\psi_h(g^j) = e^{2\pi i hj/(q-1)} \qquad \text{for } j = 0, 1, \dots, q-2$$

defines a multiplicative character of $\mathbb{F}_q$, and every multiplicative character of $\mathbb{F}_q$ is obtained in this way.

Proof. This follows immediately from Example 1.3.31. □

Remark 1.4.53 If q is a power of an odd prime, then the multiplicative character ψ_h of $\mathbb{F}_q$ in Theorem 1.4.52 with $h = (q-1)/2$ is the *quadratic character* η of $\mathbb{F}_q$. Note that for $a \in \mathbb{F}_q^*$ we have $\eta(a) = 1$ if a is the square of an element of $\mathbb{F}_q^*$ and $\eta(a) = -1$ otherwise. It is sometimes convenient to put $\eta(a) = 0$ for $a = 0 \in \mathbb{F}_q$. If q is an odd prime number p, then η agrees with the Legendre symbol for the modulus p (see Definition 1.2.22), that is, $\eta(a) = \left(\frac{a}{p}\right)$ for $a \in Z_p = \mathbb{F}_p$.

If your appetite for results on finite fields is not yet stilled, then you will find a lot of food for thought in the textbooks [74] and [103] and in the encyclopedic monographs [102] and [181]. The *Handbook of Finite Fields* edited by Mullen and Panario [118] contains over 80 survey articles on all imaginable aspects of finite fields. The aficionados of finite fields will encounter many applications of these beautiful structures in the present book.

Exercises

1.1 For all nonzero integers a and b, prove that

$$\gcd(a,b)\,\mathrm{lcm}(a,b) = |ab|.$$

1.2 For all nonzero integers a, b, c, prove that

$$\gcd(ab, ac, bc)\,\mathrm{lcm}(a,b,c) = |abc|.$$

1.3 Given $k \geq 2$ integers $a_1, \dots, a_k$ that are not all 0, prove that there exist integers $b_1, \dots, b_k$ such that

$$\gcd(a_1, \dots, a_k) = \sum_{j=1}^{k} a_j b_j.$$

1.4 Let $a, b, k \in \mathbb{N}$ with $k \geq 2$ be such that $\gcd(a,b) = 1$ and ab is a kth power of a positive integer. Prove that a and b are kth powers of positive integers.

1.5 Show that $n! + 1$ and $(n + 1)! + 1$ are coprime for all $n \in \mathbb{N}$.

1.6 Prove that the product of any four consecutive integers is divisible by 24.

1.7 Modify Euclid's trick in the proof of Theorem 1.1.12 in order to prove that there are infinitely many prime numbers that are congruent to 3 modulo 4.

1.8 Compute the least residue of 2^{34} modulo 5.

1.9 The *Euclidean algorithm* for the computation of $\gcd(a, b)$ for $a, b \in \mathbb{N}$ proceeds as follows. We can assume that $a > b$ and that b does not divide a. Then we carry out repeated divisions with remainder: $a = q_1 b + r_1$ $(1 \le r_1 < b)$, $b = q_2 r_1 + r_2$ $(0 \le r_2 < r_1)$, $r_1 = q_3 r_2 + r_3$ $(0 \le r_3 < r_2)$, and so on. Prove that this algorithm terminates after finitely many steps and that the last nonzero remainder r_j is equal to $\gcd(a, b)$. (Hint: show by induction that $\gcd(a, b) = \gcd(r_{i-1}, r_i)$ for $1 \le i \le j$, where $r_0 := b$.)

1.10 Prove that the number of steps in the Euclidean algorithm in the preceding exercise is at most $C \log b$ with an absolute constant $C > 0$.

1.11 Compute $\gcd(123, 45)$ by the Euclidean algorithm in Exercise 1.9.

1.12 For $m, b \in \mathbb{N}$ with $\gcd(m, b) = 1$, the Euclidean algorithm (see Exercise 1.9) for the computation of $\gcd(m, b)$ can be used to find an integer c with $bc \equiv 1 \pmod{m}$. We can assume that $1 < b < m$. Now start from the identity $r_{j-2} = q_j r_{j-1} + r_j$ (with $r_{-1} := m$ if $j = 1$) and note that $r_j = 1$. Hence $1 = r_{j-2} - q_j r_{j-1}$. Then show that we can run backwards through the Euclidean algorithm until we get 1 as a linear combination of b and m.

1.13 Use the method in Exercise 1.12 to determine the unique integer $c \in Z_{97}$ for which $36c \equiv 1 \pmod{97}$.

1.14 Let $a, b, m \in \mathbb{Z}$ with $m \ge 1$. Prove that there exists an integer c with $ac \equiv b \pmod{m}$ if and only if $\gcd(a, m)$ divides b.

1.15 Prove that if $ab \equiv ac \pmod{m}$ with $a, b, c \in \mathbb{Z}$, $m \in \mathbb{N}$, and $\gcd(a, m) = d$, then $b \equiv c \pmod{m/d}$.

1.16 Prove that if $a \equiv b \pmod{m}$ for $a, b \in \mathbb{Z}$ and $m \in \mathbb{N}$, then $\gcd(a, m) = \gcd(b, m)$.

1.17 If $a, b \in \mathbb{Z}$ with $a \ge 0$ and $b \ge 3$, prove that $2^a + 1$ is not divisible by $2^b - 1$.

1.18 Prove that there is no right triangle with all side lengths being integers and such that the lengths of the two sides forming the right angle are odd.

1.19 For odd $m \in \mathbb{N}$, prove that the sum of the elements of every complete residue system modulo m is divisible by m.

1.20 For $m, n \in \mathbb{N}$ with $\gcd(m, n) = 1$, prove that Euler's totient function ϕ satisfies $\phi(mn) = \phi(m)\phi(n)$.

1.21 For $m, n \in \mathbb{N}$ with $\gcd(m, n) > 1$, prove that $\phi(mn) > \phi(m)\phi(n)$.

1.22 Let $m, n \in \mathbb{N}$ be such that every prime factor of m is also a prime factor of n. Prove that $\phi(mn) = m\phi(n)$.

1.23 Prove that $\phi(m)$ is even for all integers $m \geq 3$.

1.24 Find the least positive integer a such that $a \equiv 4 \pmod{7}$, $a \equiv 2 \pmod{11}$, and $a \equiv 11 \pmod{13}$.

1.25 Find all quadratic residues modulo 13 in the least residue system modulo 13.

1.26 For every prime number $p \geq 5$, prove that the sum of the quadratic residues modulo p in any complete residue system modulo p is divisible by p.

1.27 Let p be an odd prime number and let a be a quadratic residue modulo p. Prove that for every $k \in \mathbb{N}$ there exists an integer b_k with $b_k^2 \equiv a \pmod{p^k}$.

1.28 For a prime number $p \neq 3$, an integer a with $\gcd(a, p) = 1$ is called a *cubic residue* modulo p if there exists an integer b such that $a \equiv b^3 \pmod{p}$. Prove that if $p \equiv 2 \pmod{3}$, then all integers coprime to p are cubic residues modulo p, whereas if $p \equiv 1 \pmod{3}$, then there are exactly $(p-1)/3$ cubic residues modulo p in the least residue system modulo p.

1.29 Let G be a finite abelian group with the multiplicative notation and let $a, b \in G$. Prove that $\operatorname{ord}(ab) = \operatorname{ord}(a)\operatorname{ord}(b)$ whenever $\operatorname{ord}(a)$ and $\operatorname{ord}(b)$ are coprime.

1.30 Prove that $\operatorname{ord}(a^{-1}) = \operatorname{ord}(a)$ for all elements a of a finite abelian group.

1.31 Prove that the finite abelian group R_m in Example 1.3.7 is not cyclic if $m = 2^k$ with an integer $k \geq 3$.

1.32 Let H be a subgroup of the finite abelian group G. Prove that there are exactly $|G|/|H|$ characters χ of G with the property that $\chi(h) = 1$ for all $h \in H$.

1.33 For characters χ and σ of the finite abelian group G of order t, prove that

$$\sum_{a \in G} \chi(a)\overline{\sigma(a)} = \begin{cases} t & \text{if } \chi = \sigma, \\ 0 & \text{if } \chi \neq \sigma, \end{cases}$$

where the bar denotes complex conjugation.

1.34 If ψ is a nontrivial multiplicative character and χ a nontrivial additive character of $\mathbb{F}_q$, then the *Gauss sum* $G(\psi, \chi)$ is defined by

$$G(\psi, \chi) = \sum_{c \in \mathbb{F}_q^*} \psi(c)\chi(c).$$

Prove that $|G(\psi, \chi)| = q^{1/2}$. (Hint: start from $|G(\psi, \chi)|^2 = G(\psi, \chi)\overline{G(\psi, \chi)}$, where the bar denotes complex conjugation.)

1.35 For a nontrivial additive character χ of $\mathbb{F}_q$ with q odd and for $a \in \mathbb{F}_q^*$ and $b \in \mathbb{F}_q$, put

$$S(\chi; a, b) = \sum_{c \in \mathbb{F}_q} \chi(ac^2 + b).$$

(a) Prove that

$$S(\chi; a, b) = \chi(b)\eta(a)G(\eta, \chi),$$

where η is the quadratic character of $\mathbb{F}_q$ in Remark 1.4.53.

(b) Deduce that $|S(\chi; a, b)| = q^{1/2}$.

(c) Prove $|S(\chi; a, b)| = q^{1/2}$ also directly without the use of Gauss sums. (Hint: start from $|S(\chi; a, b)|^2 = S(\chi; a, b)\overline{S(\chi; a, b)}$, where the bar denotes complex conjugation.)

1.36 For nontrivial multiplicative characters ψ and σ of $\mathbb{F}_q$, the *Jacobi sum* $J(\psi, \sigma)$ is defined by

$$J(\psi, \sigma) = \sum_{c \in \mathbb{F}_q \setminus \{0,1\}} \psi(c)\sigma(1 - c).$$

Prove that if $\psi\sigma$ is also a nontrivial multiplicative character of $\mathbb{F}_q$, then

$$J(\psi, \sigma) = \frac{G(\psi, \chi)G(\sigma, \chi)}{G(\psi\sigma, \chi)},$$

where χ is any nontrivial additive character of $\mathbb{F}_q$. (Hint: start from the product $G(\psi, \chi)G(\sigma, \chi)$ of Gauss sums.)

1.37 Prove that if ψ, σ, and $\psi\sigma$ are nontrivial multiplicative characters of $\mathbb{F}_q$, then the Jacobi sum $J(\psi, \sigma)$ satisfies $|J(\psi, \sigma)| = q^{1/2}$.

1.38 Let ψ be a nontrivial multiplicative character of $\mathbb{F}_q$ and let S be a subset of $\mathbb{F}_q$ with $h \geq 1$ elements. Prove that

$$\sum_{c \in \mathbb{F}_q} \Big| \sum_{a \in S} \psi(c + a) \Big|^2 = h(q - h),$$

where we put $\psi(0) = 0$.

1.39 Prove Theorem 1.4.22 in detail.

1.40 Let $f(x)$ and $g(x)$ be monic polynomials over an arbitrary field. Prove that

$$\gcd(f(x), g(x)) \operatorname{lcm}(f(x), g(x)) = f(x)g(x).$$

1.41 Let $f_1(x), \ldots, f_k(x)$ be $k \geq 2$ monic polynomials over an arbitrary field that are pairwise coprime. Prove that

$$\operatorname{lcm}(f_1(x), \ldots, f_k(x)) = f_1(x) \cdots f_k(x).$$

1.42 Let F be a field and let $f(x), g(x), m(x) \in F[x]$ with $m(x) \neq 0 \in F[x]$. Prove that the congruence $f(x)h(x) \equiv g(x) \pmod{m(x)}$ has a solution $h(x) \in F[x]$ if and only if $\gcd(f(x), m(x))$ divides $g(x)$ in $F[x]$.

1.43 Consider the polynomial ring $F[x]$ for an arbitrary field F. Prove the Chinese remainder theorem for $F[x]$: if $k \geq 2$ pairwise coprime nonzero polynomials $m_1(x), \ldots, m_k(x) \in F[x]$ and arbitrary polynomials $f_1(x), \ldots, f_k(x) \in F[x]$ are given, then there exists a polynomial $g(x) \in F[x]$ with $g(x) \equiv f_j(x) \pmod{m_j(x)}$ for $1 \leq j \leq k$ and $g(x)$ is uniquely determined modulo $m_1(x) \cdots m_k(x)$.

1.44 Prove the product rule for the derivative of polynomials over an arbitrary field.

1.45 Prove in detail that there are exactly $\phi(q-1)$ primitive elements in every finite field $\mathbb{F}_q$.

1.46 Set up addition and multiplication tables for the finite field $\mathbb{F}_9$.

1.47 Determine all primitive elements of $\mathbb{F}_9$.

1.48 Prove that if p is a prime number and $n \in \mathbb{N}$, then n divides $\phi(p^n - 1)$. (Hint: consider the primitive elements of the finite field $\mathbb{F}_{p^n}$.)

1.49 Prove that for $q \geq 3$, the sum of all elements of $\mathbb{F}_q$ is equal to 0.

1.50 Prove that $x^2 + x + 4 \in \mathbb{F}_{11}[x]$ is irreducible over $\mathbb{F}_{11}$.

1.51 Find all irreducible polynomials over $\mathbb{F}_2$ of degree 4.

1.52 Let $\mathbb{F}_q$ be a finite field of characteristic p. Prove that the derivative $f'(x)$ of $f(x) \in \mathbb{F}_q[x]$ is the zero polynomial if and only if $f(x)$ is the pth power of some polynomial in $\mathbb{F}_q[x]$.

1.53 Determine the minimal polynomial of $\alpha = (1 + \sqrt{5})/2$ over $\mathbb{Q}$.

1.54 For $F = \mathbb{F}_q$ and $K = \mathbb{F}_{q^n}$ with $n \in \mathbb{N}$, prove that $\mathrm{Tr}_{K/F}(\alpha^q) = \mathrm{Tr}_{K/F}(\alpha)$ for all $\alpha \in K$.

1.55 Let K be a finite extension field of the finite field F of characteristic p. Prove that

$$\mathrm{Tr}_{K/F}(\alpha^{p^n}) = (\mathrm{Tr}_{K/F}(\alpha))^{p^n} \qquad \text{for all } \alpha \in K \text{ and } n \in \mathbb{N}.$$

1.56 Prove the transitivity of the trace, that is, if $F \subseteq K \subseteq E$ are finite fields, then

$$\mathrm{Tr}_{E/F}(\alpha) = \mathrm{Tr}_{K/F}\left(\mathrm{Tr}_{E/K}(\alpha)\right) \qquad \text{for all } \alpha \in E.$$

1.57 Let α be algebraic over $F = \mathbb{F}_q$, let $m(x) \in \mathbb{F}_q[x]$ be the minimal polynomial of α over $\mathbb{F}_q$, and suppose that $\deg(m(x)) = n$. Then show that $\mathrm{Tr}_{K/F}(\alpha) = -c_{n-1}$ with $K = \mathbb{F}_{q^n}$, where c_{n-1} is the coefficient of x^{n-1} in $m(x)$.

1.58 Let K be a finite extension field of $F = \mathbb{F}_q$ and let $\alpha \in K$. Prove that $\mathrm{Tr}_{K/F}(\alpha) = 0$ if and only if $\alpha = \beta^q - \beta$ for some $\beta \in K$.

Chapter 2

Cryptography

Don't shed any tears for Bob,
this blundering bungling slob.
He mixed up n, p, and q,
giving RSA hackers a clue,
no wonder he's lynched by a mob.

2.1 Classical cryptosystems

2.1.1 Basic principles

Cryptology in the modern sense is the theory of data security and data integrity. Cryptology as a practical craft can be traced back several thousand years as it was already used in one form or other in the ancient civilizations of Egypt, Mesopotamia, China, Greece, and Rome. It would lead us too far astray if we were to delineate the colorful history of cryptology here, but we will occasionally mention some tidbits. A systematic account of the history of cryptology up to 1967 is given in the book of Kahn [75]. The more recent treatment by Singh [187] offers very stimulating reading.

Cryptology splits up into cryptography, that is, the design of secure data and communication systems, and cryptanalysis, that is, the breaking of such systems. Cryptanalysis is slippery territory: if we provide too much information here, we will be accused of giving a tutorial on hacking and cybercrime. Therefore we focus on cryptography and discuss only in general terms what is involved in cracking certain cryptographic schemes. Cryptography has various important facets, such as confidentiality (guaranteeing that sensitive messages cannot be read by eavesdroppers), data integrity (guaranteeing that the contents of messages cannot be tampered with), authentication (proving the identity of legitimate users), and non-repudiation (guaranteeing that actions such as sending messages and signing electronic documents cannot be denied later). In this chapter, we examine those aspects of cryptography where number theory plays a significant role. It would be tempting to treat also some curious angles of cryptography like hiding secret messages in pictures (a technique that is called steganography) or in poems (no special designation here, so we offer cryptopoetry), but we guess we have to show some restraint.

Let us first of all address a primary concern of cryptography, namely the protection of confidential communication. We use the technical term *channel* for a communication medium. A channel can, for instance, be a computer network, a satellite link, the Internet, or a telephone line. An important player in our scenario is the adversary, who also goes by other terms of endearment like the opponent, the enemy, the eavesdropper, the attacker, and the bad guy. The adversary wants to overhear our confidential communication and/or steal our sensitive data. Adversaries can, for instance, be your boss, the NSA, a hacker, or the parents of your lover. We speak of an *insecure channel* if we want to signalize that there is an adversary lurking around the channel.

The basic tool for the protection of confidential communication is *encryption*, that is, the transformation of given data (or messages) into disguised data (or messages) that do not give any clue about the original meaning. The reverse process of recovering the original data/messages from the encrypted data/messages is called *decryption*. A secure communication system is described by the following model.

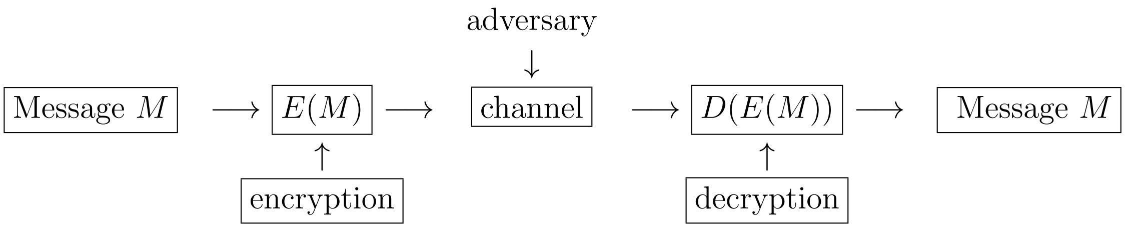

Figure 2.1: A secure communication system

The message in its original form, also called the *plaintext*, is first encrypted before it is sent over the insecure channel. Thus, the sender takes the plaintext M, applies an *encryption function* (or an *encryption algorithm*) E to it, and then transmits the *ciphertext* $C = E(M)$ over the channel. Upon receipt of C, the receiver decrypts the ciphertext C by computing $D(C) = D(E(M)) = M$ with the help of a *decryption function* (or a *decryption algorithm*) D, thus recovering the plaintext M. In this way, the secure communication is completed. Obviously, the encryption function E must be injective so that there is no ambiguity about recovering the original message correctly. Note that E and D are inverse functions of each other. Often, E and D belong to a parametrized family of functions. The parameters K and K' specifying E and D, respectively, are called the *encryption key* and the *decryption key*, respectively. Sometimes one writes E_K and $D_{K'}$ instead of E and D for the sake of clarity.

It is a generous and also prudent assumption in cryptology that the adversary has information about the general form of the encryption and decryption algorithms and can access the ciphertext. Cynics like to say that the hacker can gain all this knowledge by charming or bribing the secretary. Consequently, the security of the system is based on carefully protecting the only data that are not assumed to be available to the adversary, namely the keys. This is neatly summarized in the so-called Kerckhoff principle: *the security resides in the secret key.* For a secure system, the set of possible keys (the *key space*) must be very large in order to prevent a brute-force attack by exhaustive search for the key. The important part in the design of a secure communication system is the choice of the encryption algorithm and of the key space.

Definition 2.1.1 A *cryptosystem* (or a *cipher*) consists of an encryption algorithm (including the encryption keys) and a decryption algorithm (including the decryption keys) together with the *plaintext source*, that is, the set of all possible plaintexts.

In early cryptography, the security of a cryptosystem was based on a key exchanged by a reliable method such as a face-to-face meeting or a dispatch via a trusted courier. In modern cryptography, for example when transferring confidential information via the Internet, such a basic and simple key exchange is usually assumed to be impracticable.

There are two fundamentally different techniques for encrypting information: symmetric encryption, also called secret-key (or private-key) encryption, and asymmetric encryption, also called public-key encryption. In a *symmetric cryptosystem* the encryption and decryption keys are identical or easy to obtain from each other. Examples are block ciphers and stream ciphers. In an *asymmetric cryptosystem* the encryption and decryption keys are hard to obtain from each other without insider knowledge. Examples are the RSA cryptosystem and the ElGamal cryptosystem.

Number theory is involved in the construction of many cryptosystems. We will present quite a few such cryptosystems in this chapter. In spite of the very long history of cryptology, the serious applications of number theory (and of mathematics in general for that matter) to this area are a very recent phenomenon. In fact, the decade of the 1970s can be pinpointed as the period when these applications began in earnest. We will elaborate on the circumstances of this remarkable development in Section 2.3.

Cryptology touches many areas, such as mathematics, information theory, computer science, electrical engineering, and espionage, and so there is wide interest in the subject. This is reflected also in the large number of textbooks that have been written on cryptology. The Renaissance scholar Johannes Trithemius was possibly the first textbook author in the history of cryptology with his *Polygraphiae* published in the early 16th century. Modern readers will probably prefer more recent offerings such as Stinson [193] and van Tilborg [194]. The books of Buchmann [15] and Koblitz [82] emphasize number-theoretic aspects of cryptography, whereas Trappe and Washington [196] cover cryptography together with coding theory. An extensive treatment of cryptography from the viewpoint of computer science is given in the monograph [160]. A milestone is the *Handbook of Applied Cryptography* edited by Menezes, van Oorschot, and Vanstone [116] which may be regarded as the encyclopedia of cryptography.

2.1.2 Substitution ciphers

When one thinks about encryption, probably the first idea that comes to mind is to encrypt a message letter by letter according to a prescribed scheme. This is what a *substitution cipher* does. Formally, the set of all possible plaintexts in a substitution cipher is the 26-letter English alphabet {A,B,C,...,X,Y,Z}. The encryption function E is a permutation of the 26 alphabetic characters. The permutation is secret, that is, it is known only to the legitimate users of the system, and it is the key of the cipher. The decryption function D is the inverse permutation of E. Messages containing more than one letter are encrypted by applying E to each individual letter of the message.

Example 2.1.2 Here is an example of a substitution cipher. The encryption function E is given by the table

A	B	C	D	E	F	G	H	I	J	K	L	M
X	N	Y	A	H	P	O	G	Z	Q	W	B	T
N	O	P	Q	R	S	T	U	V	W	X	Y	Z
S	F	L	R	C	V	M	U	E	K	J	D	I

and the decryption function D is given by the table

A	B	C	D	E	F	G	H	I	J	K	L	M
D	L	R	Y	V	O	H	E	Z	X	W	P	T
N	O	P	Q	R	S	T	U	V	W	X	Y	Z
B	G	F	J	Q	N	M	U	S	K	A	C	I

The encryption and decryption algorithms are permutations of the 26 letters and the keys (here the tables of function values) specify the permutations. This substitution cipher encrypts the message cum desperate plea DAD SEND MONEY into AXA VHSA TFSHD.

Remark 2.1.3 There are altogether 26! ($\approx 4 \cdot 10^{26}$) permutations of the 26 alphabetic characters. Thus, the key space has size $\approx 4 \cdot 10^{26}$. This is too large to find the correct key by hand. However, if we use a powerful computer, then it is quite manageable to determine the key, particularly if we use additional information such as the fact that the letters of the alphabet appear with different frequencies in natural languages. For instance, in English the most frequent letter is E, the second most frequent letter is T, and so on. Therefore, if in a long ciphertext from a substitution cipher the letter Q occurs most frequently, then we can deduce with good confidence that Q is the encryption of E, and similarly for other frequent letters. The original message can then be recovered by using the redundancy of natural languages and some combinatorial skill. Because of this vulnerability, substitution ciphers are only of historical and didactic interest.

The *affine cipher* is a special case of the substitution cipher and it is based on modular arithmetic with the modulus $m = 26$.

Algorithm 2.1.4 (Affine Cipher) The plaintext source is the set of the 26 English letters, identified with the least residue system Z_{26} modulo 26 via the correspondence

$$A \leftrightarrow 0, B \leftrightarrow 1, \ldots, Z \leftrightarrow 25.$$

The encryption key consists of a positive integer a with $\gcd(a, 26) = 1$ and an integer b, both considered modulo 26, and the encryption algorithm is given by the function $e : Z_{26} \to Z_{26}$ determined by $e(m) \equiv am + b \pmod{26}$ for all $m \in Z_{26}$. The decryption key consists of the integer $c \in Z_{26}$ with $ac \equiv 1 \pmod{26}$ and b, and the decryption algorithm is given by the function $d : Z_{26} \to Z_{26}$ with $d(r) \equiv c(r - b) \pmod{26}$ for all $r \in Z_{26}$.

Remark 2.1.5 Here we clearly see the general structure of a cryptosystem described in Subsection 2.1.1. The encryption and decryption functions belong to the family of affine functions modulo 26 and the keys $K = (a, b)$ and $K' = (c, b)$ specify these affine functions.

Remark 2.1.6 Affine ciphers with $a = 1$ are called *shift ciphers* since the encryption function is then a cyclic shift in the alphabet by b letters. The choice $a = 1$ and $b = 3$ yields the *Caesar cipher* which was used by the Roman emperor Julius Caesar according to his biographer Suetonius. Many scholars regard this as the first concretely and explicitly recorded cipher in the history of cryptology, while vague references to shift ciphers can be found also in earlier documents. (The first author named his first dog Caesar as a tribute to the pioneer cryptographer Julius Caesar.)

Remark 2.1.7 There are movie fans who believe that the name of the computer HAL in the Stanley Kubrick film *2001: A Space Odyssey* was obtained by using a shift cipher. Indeed, if you take the shift cipher with $b = 25$, which is a shift backward by one letter in the alphabet, and apply it to IBM, then *voilà* you get HAL.

For a given encryption key $K = (a, b)$ in an affine cipher, determining the decryption key amounts to finding the integer $c \in Z_{26}$ with $ac \equiv 1 \pmod{26}$ as specified in Algorithm 2.1.4. This can be easily done by using Euler's theorem (see Theorem 1.2.15) and the square-and-multiply algorithm described in Subsection 2.3.2 below. Alternatively and somewhat more efficiently, we can use the Euclidean algorithm which is discussed in [152, Section 1.2] (see also Exercise 1.12). Thus, the affine cipher is a symmetric cryptosystem.

Remark 2.1.8 Since the affine cipher has a very small key space of size $\phi(26) \cdot 26 = 312$, it can be broken easily, with some patience even by hand, using exhaustive key search.

Remark 2.1.9 The substitution and affine ciphers are called *monoalphabetic ciphers* because every single alphabetic character is mapped to a unique alphabetic character for a fixed encryption key. If we collect blocks of $n \geq 2$ letters together and encrypt each such block, then we get a *polyalphabetic cipher*. A well-known historical example of a polyalphabetic cipher dating back to the 16th century is the *Vigenère cipher* in which n shift ciphers are applied in parallel and the keys for these n shift ciphers are chosen independently of each other. Therefore the size of the key space is 26^n. This is definitely an improvement on the affine cipher, and we can indeed get very large key spaces by this method if we are willing to make n large.

Polyalphabetic ciphers were mechanized in the 20th century, and a cipher machine that became famous was the Enigma. The Enigma cipher and its implementation on the Enigma machine were used by the German military from the 1920s on. During World War II the Allies raised a monumental effort to crack the Enigma cipher, based on earlier progress made by Polish cryptanalysts, and they succeeded fairly early in the war. Some historians claim that this achievement had a major impact on the outcome of the war. The breaking of the Enigma cipher is a story full of drama and suspense, and so it is not surprising that two blockbuster movies were made on this subject: *Enigma* starring Kate Winslet and Dougray Scott and *The Imitation Game* starring Keira Knightley as well as Benedict Cumberbatch in the role of the brilliant mathematician and pioneer computer scientist Alan Turing (1912–1954) who was a decisive factor in the cracking of Enigma.

2.2 Symmetric block ciphers

2.2.1 Data Encryption Standard (DES)

Now we move from historical cryptosystems to a family of ciphers that are widely utilized in our present age, namely symmetric block ciphers. This family includes the industry standards DES and AES. Cryptochips running the DES algorithm or the AES algorithm are omnipresent in the automatic teller machines (ATMs) that supply us with our daily cash. The point is that the communication between an ATM and the server of the bank is highly confidential and therefore has to be protected by encryption. The encryption algorithm must be able to process large amounts of data at a very high speed, and this is where DES and AES shine.

Definition 2.2.1 A *block cipher* splits up the plaintext into blocks of symbols of fixed length (for example n bits) and encrypts each block in a manner that is independent of past input blocks. The ciphertext depends only on the current input block and on the key. In a *symmetric block cipher*, the encryption and decryption keys are identical or easy to obtain from each other, and they are kept secret.

In every practical implementation of a symmetric block cipher, the plaintext is given as a string of bits. If the plaintext has some other format, it first has to be transformed into a string of bits before applying a symmetric block cipher.

Example 2.2.2 The Vigenère cipher in Remark 2.1.9 can be viewed as a bit-based symmetric block cipher if letters are transformed into blocks of bits by using for example the ASCII code.

The *Data Encryption Standard* (*DES*) is a symmetric block cipher that was developed by IBM and endorsed by the U.S. National Bureau of Standards in 1977. It has been widely used ever since, for example in the banking industry as mentioned above, and its design details are available to the public.

DES encrypts plaintext blocks of 64 bits. The user first chooses a key consisting of 56 random bits, which is then split into eight blocks of seven bits each. For error control (compare with Section 6.1), a parity-check bit is added to each block of seven bits, that is, the check bit is 0 or 1 depending on whether the number of 1's in the previous seven bits is even or odd, respectively. Thus, the actual key K has length 64, but only 56 bits (now in positions $1, 2, \ldots, 7, 9, \ldots, 15, \ldots, 57, \ldots, 63$) have been chosen by the user. The same key is applied in both the encryption and the decryption and is of course kept secret. The effective size of the key space is 2^{56}.

Let $M = m_1 m_2 \ldots m_{64}$ be a given plaintext block of 64 bits m_j, $1 \leq j \leq 64$. The DES encryption algorithm first applies a fixed permutation P to M, namely

$$P(M) = m_{\pi(1)} m_{\pi(2)} \ldots m_{\pi(64)}$$

with a permutation π of $\{1, \ldots, 64\}$ given by

$$\pi(i) \equiv \begin{cases} 58i \pmod{66} & \text{for } i = 1, 2, \ldots, 32, \\ 58(i-32) - 1 \pmod{66} & \text{for } i = 33, 34, \ldots, 64. \end{cases}$$

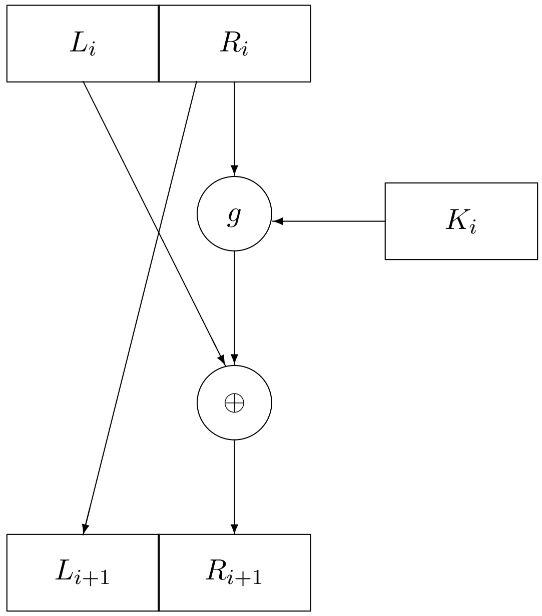

Figure 2.2: A typical round of DES

Then 16 iterations of a function f are applied to $P(M)$. Finally, the inverse permutation P^{-1} operates on the last output, and this produces the ciphertext. Note that for $i = 1, \ldots, 64$, the image $\pi^{-1}(i)$ under the inverse permutation π^{-1} of π is

$$\pi^{-1}(i) = \begin{cases} r_{33}(4i) & \text{if } i \text{ is even,} \\ r_{33}(4i+3) + 33 & \text{if } i \text{ is odd,} \end{cases}$$

where $r_{33}(a)$ denotes the least residue of the integer a modulo 33. The function f combines substitution and transposition. A typical operation in the calculation of f proceeds as follows. If $T_i = t_1 t_2 \ldots t_{64}$ (with bits t_j for $1 \le j \le 64$) is an intermediate result, then split up T_i into the left half L_i and the right half R_i, that is, $T_i = L_i R_i$ with

$$L_i = t_1 \ldots t_{32}, \qquad R_i = t_{33} \ldots t_{64}.$$

Then

$$L_{i+1} = R_i, \qquad R_{i+1} = L_i \oplus g(R_i, K_i),$$

where g is a known function with range $\mathbb{F}_2^{32}$ and K_i is an intermediate key derived from the key K. Here $\oplus$ denotes bit by bit addition in $\mathbb{F}_2$ of two blocks of 32 bits. For the full details of the algorithm we refer to [116, Section 7.4]. Decryption is performed by essentially the same algorithm, except that the order of the intermediate keys is reversed.

Over time, several weaknesses of DES were discovered by cryptanalysts. The key of 56 bits is now considered too short. A better variant still in current use is *Triple DES*. Here a plaintext block M is encrypted as

$$C = \mathrm{DES}_{K_1}(\mathrm{DES}_{K_2}^{-1}(\mathrm{DES}_{K_1}(M)))$$

b_{00}	b_{01}	b_{02}	b_{03}
b_{10}	b_{11}	b_{12}	b_{13}
b_{20}	b_{21}	b_{22}	b_{23}
b_{30}	b_{31}	b_{32}	b_{33}

Figure 2.3: A 4×4 array representing a plaintext block. Each b_{ij} is a byte and also an element of $\mathbb{F}_{256}$.

with an obvious notation, that is, M is encrypted, decrypted, and then encrypted again by DES, using a key K_1 for the encryptions and a different and independently chosen key K_2 for the decryption. The effective size of the key space is now 2^{112} and this yields a higher security level compared to DES.

2.2.2 Advanced Encryption Standard (AES)

The increasing dissatisfaction with DES, which stood more and more for Deficient Encryption Scheme, called for remedial action. The U.S. National Institute of Standards and Technology (NIST) ran a competition to find a state-of-the-art symmetric block cipher succeeding DES. Submissions were received in 1998 and 15 of them met the criteria of NIST. Five finalists were selected in 1999 and the winner was announced in 2000. The winning design was from Belgium and was called Rijndael at the time of submission, after the names of the designers Rijmen and Daemen. It seems that Rijndael is also a play on words in the Dutch language since *Rijndal* means Rhine valley in Dutch. Rijndael became the *Advanced Encryption Standard* (*AES*) in 2001. There are actually slight design differences between Rijndael and the officially adopted Advanced Encryption Standard, but only AES will be discussed here.

The plaintext block length in AES is 128 bits and the key length can be 128, 192, or 256 bits. Many operations in AES are based on bytes. A *byte*, that is, a string $a_0a_1 \dots a_7$ of eight bits, is identified with the polynomial $a_0+a_1x+\cdots+a_7x^7 \in \mathbb{F}_2[x]$. This polynomial is in turn interpreted as an element of the finite field $\mathbb{F}_{256}$ which is viewed as the residue class field $\mathbb{F}_2[x]/(x^8+x^4+x^3+x+1)$. Note that the polynomial $x^8+x^4+x^3+x+1 \in \mathbb{F}_2[x]$ is irreducible over $\mathbb{F}_2$, and so the residue class ring $\mathbb{F}_2[x]/(x^8+x^4+x^3+x+1)$ is indeed a finite field of order $2^8 = 256$ (see Subsection 1.4.3).

A given plaintext block of 128 bits is split up into 16 blocks of eight bits, that is, into 16 bytes. Each byte is interpreted as an element of $\mathbb{F}_{256}$ as above. The 16 bytes are arranged into a 4×4 array. Thus, the plaintext block is finally viewed as a 4×4 array of elements of $\mathbb{F}_{256}$. The rows and columns of the array are indexed by $0, 1, 2, 3$.

The AES encryption algorithm has a number of rounds, each consisting of four operations: SubBytes, ShiftRows, MixColumns, and AddRoundKey. In the standard case of key length 128, the number of rounds is 10.

SubBytes has two steps: (i) each array element is replaced by its multiplicative inverse in $\mathbb{F}_{256}$, with 0 being mapped to 0; (ii) the array undergoes a fixed affine transformation over $\mathbb{F}_{256}$.

ShiftRows cyclically shifts the elements of the ith row ($i = 0, 1, 2, 3$) of the array by i elements to the left.

MixColumns views each column of the array as a polynomial over $\mathbb{F}_{256}$ of degree at most 3 and multiplies this polynomial by a fixed polynomial over $\mathbb{F}_{256}$ that is coprime to $x^4+1 \in \mathbb{F}_2[x]$, with reduction modulo x^4+1 in the case of overflow. The new polynomial yields the new column.

AddRoundKey adds the array bit by bit, using addition in $\mathbb{F}_2$, to another array of the same format, where the latter array depends on an intermediate key.

Decryption in AES is carried out by using the inverses of the steps and reversing their order. Note that this is feasible since each step represents an injective map. All you ever want to know about Rijndael and AES can be found in the book [34] which comes straight from the horse's mouth, namely from the designers themselves.

Being number theorists, we like to think of AES as a smarter cipher than DES since it relies much more heavily on number theory than DES does. It remains to be seen how long AES will survive. It is a known phenomenon that once a cryptographic scheme is elevated to the rank of a standard, then hordes of cryptanalysts assail it and seek the fame that is gained by breaking it.

Remark 2.2.3 Consider the permutation f of the finite field $\mathbb{F}_{2^r}$ with $r \in \mathbb{N}$ that is defined by $f(y) = y^{-1}$ for $y \in \mathbb{F}_{2^r}^*$ and $f(0) = 0$. This function is used in the first step of SubBytes with the choice $r = 8$. The function f has the following property: for all $a \in \mathbb{F}_{2^r}^*$ and $b \in \mathbb{F}_{2^r}$, the equation

$$f(y) + f(y+a) = b$$

has at most two solutions $y \in \mathbb{F}_{2^r}$ if r is odd. (Note that $y + a$ is a solution whenever y is a solution.) Such a function is called *almost perfect nonlinear* (*APN*). APN functions are the functions that best resist the so-called differential attacks on cryptosystems, see for example [19]. If r is even and $b = a^{-1}$, then the equation above has four solutions $y = 0, a, ca, (c+1)a$, where $c \in \mathbb{F}_4 \subseteq \mathbb{F}_{2^r}$ is a root of the polynomial $x^2 + x + 1 \in \mathbb{F}_2[x]$.

2.3 Public-key cryptosystems

2.3.1 Background and basics

All the cryptosystems we have discussed so far satisfy the property that the decryption key is the same as the encryption key or is easily derived from the encryption key. In other words, they are all symmetric cryptosystems. But in many modern communication systems, the big difficulty with symmetric cryptosystems is how to get the common key from user A to user B if A and B can communicate only over an insecure channel, for example by email over the Internet. This problem is solved by using public-key cryptosystems.

The idea of a public-key cryptosystem goes back to the paper of Diffie and Hellman [40] from 1976 with the prophetic title "New directions in cryptography". Their fundamental insight was that a secret key is needed only for decryption! Thus, the roles of the encryption and decryption keys can be separated: use a public key for encryption and a private (secret) key for decryption. Of course, the public key and the private key must be completely different (this is why we speak also of an *asymmetric cryptosystem*). In a nutshell, a public-key cryptosystem is built on the following new principle: *anybody can encrypt, but only the legitimate receiver can decrypt.*

The paper of Diffie and Hellman was a watershed in the history of cryptology, or a paradigm change to use a fancy term from the philosophy of science. It is generally acknowledged that cryptology as a serious mathematical discipline began only in 1976, while before that it was more like an art or a craft. The Diffie-Hellman paper triggered a burst of creativity that led to the design of many cryptographic schemes in the late 1970*s* and throughout the 1980s. Interestingly enough, many pure mathematicians entered the game of inventing cryptographic schemes during that period. Several of these schemes are still in practical use today. It is a curious footnote to the history of cryptology that the priority for the invention of public-key cryptography was later claimed by the British secret service, but they can of course assert anything under their shroud of secrecy.

The personal story behind the Diffie-Hellman paper is quite remarkable. Whitfield Diffie came to Stanford University in the mid 1970s as a somewhat overripe (or should we say mature?) graduate student (he was born in 1944). He was an autodidact in the field of cryptology since this subject was not taught at universities at that time. He met a congenial partner in Martin Hellman who was then an assistant professor at Stanford and actually younger than Diffie. Formally, Hellman was the Ph.D. adviser of Diffie, but their style of work was collaboration rather than a professor-student relationship. Their story is told in fascinating detail in the book of Levy [100] which reads like a thriller.

For a public-key cryptosystem, the encryption algorithm E and the decryption algorithm D should satisfy the following properties.

PKC1: The encryption and decryption algorithms are fast.
PKC2: $D(E(M)) = M$ for all plaintexts M in the plaintext source.
PKC3: Given the encryption key, it must be computationally infeasible to determine the decryption key.

We can set up a public-key cryptosystem in the following way. A typical user A chooses an encryption key $K_{\rm A}$ and the corresponding decryption key $K'_{\rm A}$. Then A makes the encryption key $K_{\rm A}$ public for all users and keeps the decryption key $K'_{\rm A}$ secret. The encryption key is called also the *public key* and the decryption key is called also the *private key*. Because of the property PKC3 of a public-key cryptosystem, other users cannot figure out the decryption key from the public key.

Suppose that a user A wants to send a confidential message M to another user B. Before we proceed any further, we personalize the scenario: user A is in reality called Alice and user B is in reality called Bob. So it is Alice who wants to send the message M to Bob. She proceeds as follows. She looks up the encryption key $K_{\rm B} = K_{\rm Bob}$ of Bob in a directory and then she encrypts the plaintext M into the ciphertext

$$C = E_{K_{\rm B}}(M).$$

The ciphertext C is sent to Bob through the insecure channel. Everyone may read C, but only Bob can decrypt the ciphertext C by calculating

$$D_{K'_{\rm B}}(C) = D_{K'_{\rm B}}(E_{K_{\rm B}}(M)) = M.$$

This completes the confidential communication. People other than Bob cannot decrypt the ciphertext because they cannot determine $K'_{\rm B}$ from $K_{\rm B}$.

Remark 2.3.1 We describe a "hardware" analog of a public-key cryptosystem which may help to better understand how public-key cryptosystems work. The principal tools of this analog are padlocks. Note that anybody can lock a padlock (just push it until the lock clicks), but that it can be opened only with the correct key. Now suppose that Alice wants to send a confidential document to Bob. She puts the document in a strongbox and then she goes to a sort of post office where padlocks of all users of the communication system are available. She locks the strongbox with Bob's padlock and posts the strongbox plus padlock. When Bob receives this delivery, he unlocks the padlock with his key and retrieves the document from the strongbox. We see again the guiding principle of public-key cryptosystems in operation: anybody can encrypt (lock the padlock), but only the legitimate receiver (Bob) can decrypt (unlock the padlock).

Public-key cryptosystems can be used not only for the communication of secret messages, but also for the distribution of keys in symmetric cryptosystems.

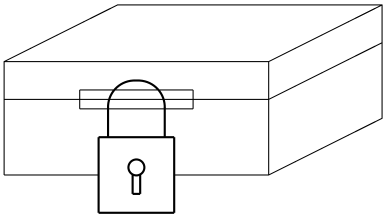

Figure 2.4: A strongbox with a padlock

Before the encryption by a symmetric cryptosystem starts, the necessary keys are distributed by a public-key cryptosystem. This makes sense because encryption by a symmetric cryptosystem is usually much faster than by a public-key cryptosystem. Note that the distribution of keys is needed only once in a communication session. Key distribution by public-key cryptosystems is a huge advantage because the parties in the communication do not even have to know each other! A scheme that uses a public-key cryptosystem for key distribution and a symmetric cryptosystem for message encryption is called a *hybrid cryptosystem*. A well-known example of a hybrid cryptosystem is Pretty Good Privacy (PGP), a popular tool for encrypting email messages.

The difficulty in designing a public-key cryptosystem is to satisfy the property PKC3 above. An important step in the design of a public-key cryptosystem is to use one-way functions as encryption functions. As usual, we write f^{-1} for the inverse function of an injective function f.

Definition 2.3.2 A *one-way function* is an injective function $f : A \to B$ with a domain A and a range B for which the following properties hold:
(i) $f(a)$ is easy to evaluate for all $a \in A$;
(ii) given f, it is computationally infeasible to compute $f^{-1}(b)$ for almost all b in the image of f.

Example 2.3.3 As a simple example from everyday life, consider the telephone directory of any big city. It is easy to look up the phone number of any specific person. On the other hand, given an arbitrarily chosen phone number, it is in general pretty hopeless to find the person with this number by just using the directory. In this sense, the function f : *person* $\mapsto$ *phone number* can be viewed as a one-way function.

Remark 2.3.4 An important application of one-way functions is the password file of a computer. It is obviously too dangerous to store passwords as plaintexts. Therefore a password P is stored as $f(P)$, where f is a one-way function. The inverse function f^{-1} is not needed in this application, for if a password P is entered, then the computer simply checks whether the image $f(P)$ coincides with the stored value. On the other hand, an intruder reading the password file has to know f^{-1} in order to deduce the password from the stored value, but f^{-1} is hard to compute.

You will concede that the definition of a one-way function is not really rigorous. What do "easy" and "computationally infeasible" mean precisely? It is generally agreed that a function evaluation is "easy" (or "efficient") if it can be carried out in polynomial time, that is, the number of required arithmetic operations (or bit operations as the case may be) is a polynomial in the number of bits of the input. "Computationally infeasible" is then somehow the opposite, but in practice it is extremely difficult to prove that it is *not* possible to compute a mathematical quantity efficiently. Furthermore, we use "almost all" in part (ii) of Definition 2.3.2 only in an intuitive sense; you may interpret it as "a very high percentage of".

Another big problem remains: if we use a one-way function f for encryption, then at least one person has to know how to invert f, namely the intended recipient of the message! The solution of this dilemma is to use *trapdoor one-way functions*, which are one-way functions such that with additional information (the *trapdoor information*) the function f can be inverted and so the ciphertext can be decrypted. The trapdoor information is known only to the authorized person who generates the encryption key and the decryption key. This authorized person can be the legitimate user who will own the keys, but also a trusted third party. In the following, we will see various ways of how to produce trapdoor information. As for many cryptographic terms, the name "trapdoor" is well chosen: it suggests a ciphertext caught helplessly in a trap, but somebody has a key for the trapdoor and frees (decrypts) the ciphertext.

2.3.2 The RSA cryptosystem

The RSA cryptosystem is named after its inventors Rivest, Shamir, and Adleman and was published in the paper [171] in 1978, hence fairly soon after the Diffie-Hellman paper [40]. The fact that a powerful and convincing public-key cryptosystem was designed shortly after Diffie and Hellman propounded their ideas gave a big boost to public-key cryptography.

The RSA cryptosystem is a public-key cryptosystem based on the presumed difficulty of finding the factorization of large integers into prime numbers. Actually, only a special case of the factorization problem is considered: given two distinct big prime numbers p and q, it is easy to find the product $n = pq$; however, given the product n, it is believed to be very hard in general to find the prime factors p and q. This belief is considered reasonable at present since up to now no deterministic polynomial-time factorization algorithm for integers is known. The situation will of course change dramatically once somebody finds such an algorithm. Consequently, there is no absolute guarantee for the security of the RSA cryptosystem, and a similar state of affairs prevails for other public-key cryptosystems.

In order to set up the RSA cryptosystem, our typical user Bob chooses two distinct big prime numbers p and q. Furthermore, Bob computes

$$n = pq \quad \text{and} \quad \phi(n) = (p-1)(q-1).$$

Then Bob chooses an integer $e \geq 2$ with $\gcd(e, \phi(n)) = 1$ and computes a positive integer d such that $ed \equiv 1 \pmod{\phi(n)}$. Note that d can be obtained efficiently by the Euclidean algorithm (see [152, Section 1.2] and Exercise 1.12).

Algorithm 2.3.5 (RSA Cryptosystem) The public key of Bob is the ordered pair (n, e) and the private key of Bob is the ordered triple (p, q, d). The plaintext source is $Z_n = \{0, 1, \ldots, n-1\}$, that is, the least residue system modulo n.
Encryption: Suppose that Alice wants to send a plaintext $m \in Z_n$ to Bob. Then Alice computes the integer $c \in Z_n$ with $c \equiv m^e \pmod{n}$ and sends c as the ciphertext to Bob.
Decryption: Upon receiving the ciphertext c, Bob computes the least residue of c^d modulo n, which is the plaintext m.

It remains to verify that the least residue of c^d modulo n is indeed the plaintext m. Note that $c^d \equiv m^{ed} \pmod{n}$, and so it suffices to prove the following lemma.

Lemma 2.3.6 *If* $ed \equiv 1 \pmod{\phi(n)}$ *as above, then*

$$m^{ed} \equiv m \pmod{n} \qquad \textit{for all } m \in \mathbb{Z}.$$

Proof. Since $ed \equiv 1 \pmod{\phi(n)}$, we can write $ed = k\phi(n) + 1$ with some positive integer k. If $\gcd(m, n) = 1$, then $m^{\phi(n)} \equiv 1 \pmod{n}$ by Theorem 1.2.15. This implies $m^{k\phi(n)} \equiv 1 \pmod{n}$, hence

$$m^{ed} \equiv m^{k\phi(n)+1} \equiv m \pmod{n}.$$

If $\gcd(m, n) = p$ or q, say p, then $m^{q-1} \equiv 1 \pmod{q}$ by Theorem 1.2.15. We infer that $m^{\phi(n)} \equiv 1 \pmod{q}$, so as above we get $m^{ed} \equiv m \pmod{q}$. But p divides m, so $m^{ed} \equiv 0 \equiv m \pmod{p}$. Together with the last congruence modulo q this yields $m^{ed} \equiv m \pmod{n}$ by the Chinese remainder theorem (see Theorem 1.2.9). The remaining case is $\gcd(m, n) = n$. But then $m^{ed} \equiv 0 \equiv m \pmod{n}$, and we have settled all cases. □

The security of the RSA cryptosystem is based on the presumed difficulty of finding d from n and e. If the opponent can factor $n = pq$, then he (the bad guy is always male) can easily compute $\phi(n) = (p-1)(q-1)$ and consequently the integer d. There is no other known efficient way of getting d than by factoring n. In this sense, the security of the RSA cryptosystem is founded on the belief that the factorization problem for large integers is difficult. The trapdoor information that allows Bob to compute d is the knowledge of the prime factors p and q. According to current standards, each of p and q should have about 150 decimal digits, so n should have about 300 decimal digits. Experts like to speak of a 1024-bit RSA modulus. The problem of finding large prime numbers will be discussed in Section 6.6. It was shown by May [112] that breaking the RSA cryptosystem is deterministic polynomial-time equivalent to factoring if p and q have the same bit size.

Remark 2.3.7 The RSA cryptosystem is so easy to understand that even criminals can use it. A psychopath terrorized Austria in the 1990s with threatening messages encrypted by the RSA cryptosystem and with letter bombs. He teased the police by including the public key (n, e) in the messages, but he made the stupid mistake of choosing prime numbers p and q that are very close together. This case is quickly broken: just compute $\lfloor \sqrt{n} \rfloor$ and look for prime numbers in the vicinity of this integer (see also Algorithm 2.3.11 below).

A practical issue that needs to be addressed in Algorithm 2.3.5 is how to compute the powers m^e and c^d modulo n for very large exponents e and d in an efficient manner. We phrase this problem in a more general form since we will run into it again in other contexts. Thus, let S be any algebraic structure in which an associative product is defined. Then for all $a \in S$ and all exponents $e \in \mathbb{N}$, the power a^n is defined unambiguously. The high-school method of computing powers by successive multiplication, that is,

$$a^e = \underbrace{a \cdot a \cdots a}_{e \text{ factors}},$$

needs $e - 1$ multiplications. This can be practically infeasible in an RSA setting where e may have several hundred decimal digits. A much faster way is provided by the *square-and-multiply algorithm*. We first explain this algorithm in an example.

Example 2.3.8 We want to compute a^{25}. We write the exponent 25 in its binary representation

$$25 = 1 + 0 \cdot 2 + 0 \cdot 2^2 + 1 \cdot 2^3 + 1 \cdot 2^4 = 1 + 8 + 16.$$

Then

$$a^{25} = a \cdot a^8 \cdot a^{16}.$$

We first calculate a^2, $a^4 = (a^2)^2$, $a^8 = (a^4)^2$, $a^{16} = (a^8)^2$ by repeated squaring. Then we multiply together a, a^8, and a^{16} to obtain a^{25}. Instead of 24 multiplications by the high-school method, we need just six multiplications by the square-and-multiply algorithm.

Algorithm 2.3.9 (Square-and-Multiply Algorithm) Let S be an algebraic structure in which an associative product is defined, let $a \in S$, and let $e \in \mathbb{N}$. Compute a^e.
Step 1: Write e in its binary representation

$$e = 2^{k_1} + 2^{k_2} + \cdots + 2^{k_r} \quad \text{with } 0 \leq k_1 < k_2 < \cdots < k_r.$$

Step 2: Compute the powers $a^2, a^4, \ldots, a^{2^{k_r}}$ by repeated squaring.
Step 3: Multiply together $a^{2^{k_1}}, a^{2^{k_2}}, \ldots, a^{2^{k_r}}$ to obtain a^e.

Proposition 2.3.10 *For all $a \in S$ and $e \in \mathbb{N}$, the computation of a^e by the square-and-multiply algorithm needs at most $2 \log_2 e$ multiplications in S, where $\log_2$ denotes the logarithm to the base 2.*

Proof. Let $h \in \mathbb{N}$ be such that $2^{h-1} \leq e < 2^h$. Then in Step 2 of Algorithm 2.3.9 we have to calculate $a^2, a^4, \ldots, a^{2^{h-1}}$. This needs $h-1$ multiplications in S. For Step 3 in Algorithm 2.3.9, in the worst case we have to multiply

$$a \cdot a^2 \cdot a^4 \cdots a^{2^{h-1}}.$$

This needs again $h-1$ multiplications in S. Altogether, we require at most $2(h-1)$ multiplications in S. The proof is completed by noting that $h - 1 \leq \log_2 e$. □

Returning to the RSA cryptosystem, we observe that even if the exponents e and d in Algorithm 2.3.5 have about 1000 bits, then encryption and decryption would each require at most about 2000 multiplications modulo n by the square-and-multiply algorithm. This is an easy task for a modern computer.

2.3.3 Factorization methods

We initially assume in this subsection that n is a product of two distinct odd prime numbers, say $n = pq$ with $p > q$. We pointed out that the RSA cryptosystem can be broken once n is factored. We note that this is equivalent to knowing the value of $\phi(n)$. First, if $n = pq$ is factored, then $\phi(n) = (p-1)(q-1)$ is obtained immediately. Conversely, if $\phi(n)$ is known, then it is easy to check that p and q are the roots of the quadratic equation

$$x^2 - (n - \phi(n) + 1)x + n = 0.$$

We now discuss some classical factorization methods. A joyful and up-to-date account of factoring is given in the book of Wagstaff [198]. A factorization algorithm using quantum computers will be presented in Subsection 6.5.1.

The first method is named after the famous 17th century mathematician Pierre de Fermat and applies if the two prime factors p and q of n are close. Then $s := (p - q)/2$ is a small number.

Algorithm 2.3.11 (Fermat Factorization) Given $n = pq$ with close prime numbers $p > q \geq 3$, we see from

$$n = \left(\frac{p+q}{2}\right)^2 - \left(\frac{p-q}{2}\right)^2$$

that $t := (p+q)/2$ is an integer slightly larger than $\sqrt{n}$ having the property that $t^2 - n = s^2$ is a perfect square. By testing the successive integers $t > \sqrt{n}$, one will soon find t and s. Then $t + s$ and $t - s$ are the two prime factors of n.

Example 2.3.12 For $n = 35$ we try $t = \lceil\sqrt{35}\rceil = 6$ and get $s = \sqrt{t^2 - n} = 1$. Hence $p = t + s = 7$ and $q = t - s = 5$.

The second method is based on the following result.

Lemma 2.3.13 *Let x, y, and n be positive integers. If $x^2 \equiv y^2 \pmod{n}$ but $x \not\equiv \pm y \pmod{n}$, then $\gcd(x-y, n)$ and $\gcd(x+y, n)$ are nontrivial divisors of n.*

Proof. Note first that n divides $x^2 - y^2 = (x-y)(x+y)$. However, n divides neither $x - y$ nor $x + y$. Hence $\gcd(n, x \pm y) > 1$. □

Example 2.3.14 For $n = 35$ we easily find the congruence $12^2 \equiv 2^2 \pmod{35}$, and then $\gcd(12-2, 35) = 5$ and $\gcd(12+2, 35) = 7$ are the prime factors of 35.

The following lemma guarantees the existence of x and y.

Lemma 2.3.15 *Let n be a product of two distinct odd prime numbers and let $a \in \mathbb{Z}$ with $\gcd(a, n) = 1$. Then $x^2 \equiv a^2 \pmod{n}$ has exactly four solutions $x \in Z_n$; two of them are the trivial solutions $x \equiv \pm a \pmod{n}$.*

Proof. For odd prime numbers p and q, the congruences $x^2 \equiv a^2 \pmod{p}$ and $x^2 \equiv a^2 \pmod{q}$ have exactly two incongruent solutions $x \equiv \pm a \pmod{p}$ and $x \equiv \pm a \pmod{q}$, respectively. If $p \neq q$, then by the Chinese remainder theorem (see Theorem 1.2.9) there are exactly four solutions of $x^2 \equiv a^2 \pmod{n}$ in Z_n with $n = pq$. □

Example 2.3.16 For $n = 35$ and $a = 2$, the congruence $x^2 \equiv 4 \pmod{35}$ has the four solutions $x \equiv \pm 2, \pm 12 \pmod{35}$ in Z_{35}, that is, $x = 2, 12, 23, 33$. The solutions $x = 2$ and $x = 33$ are the trivial ones.

The crucial step for this method is to find a nontrivial solution of $x^2 \equiv y^2 \pmod{n}$. We describe one such method.

Algorithm 2.3.17 (Square-Root Factoring) Given $n = pq$ as above, find a nontrivial solution of $x^2 \equiv y^2 \pmod{n}$.
Step 1: Choose $S = \{p_1, p_2, \dots, p_k\}$, where p_j is the jth prime in the natural order, and select an integer c slightly larger than k.
Step 2(a): For $i = 1, 2, \dots, c$ choose (randomly) $a_i \in \{0, 1. \dots, n-1\}$ and calculate $b_i \equiv a_i^2 \pmod{n}$.
Step 2(b): Write (if possible)

$$b_i = \prod_{j=1}^{k} p_j^{e_{ij}} \qquad \text{with integers } e_{ij} \geq 0.$$

Otherwise choose a new a_i.
Step 3: Find a set $T \subseteq \{1, \dots, c\}$ such that $\prod_{i \in T} b_i$ is a square by determining a nontrivial linear combination over $\mathbb{F}_2$ of the c binary vectors $\mathbf{v}_i \equiv (e_{i1}, \dots, e_{ik}) \pmod{2}$, $1 \leq i \leq c$, which yields $\mathbf{0} \in \mathbb{F}_2^k$.
Step 4: Compute $x = \prod_{i \in T} a_i$ and $y = \left(\prod_{i \in T} b_i\right)^{1/2}$.
Step 5: If $x \not\equiv \pm y \pmod{n}$, then stop with success. If $x \equiv \pm y \pmod{n}$, then take a different set T or increase the value of c by 1.

The following two algorithms are due to the prolific algorithm designer John Pollard and they apply to any composite integer n. Let B be a positive integer. An integer $n \geq 2$ is said to be B-*smooth* if all its prime factors are less than or equal to B.

Algorithm 2.3.18 (Pollard $p - 1$ Algorithm) Find a nontrivial factor of the composite integer n.
Step 1: Select a smoothness bound B.
Step 2: Select a random integer a with $2 \leq a \leq n - 1$ and compute $d = \gcd(a, n)$. If $d \geq 2$, then return the nontrivial factor d of n.
Step 3: If $d = 1$, then for each prime number $r \leq B$ perform the following iteration: compute $l(r) = \lfloor (\log n)/\log r \rfloor$ and replace the current value of a by the least residue of $a^{r^{l(r)}}$ modulo n (use Algorithm 2.3.9 to compute this least residue efficiently).
Step 4: Compute $e = \gcd(b - 1, n)$, where b is the last output in Step 3.
Step 5: If $1 < e < n$, then e is a nontrivial factor of n. Otherwise return to Step 2 and choose another integer a.

Note that in Step 3 we are calculating the least residue b of a^Q modulo n, where

$$Q = \prod_{r \leq B} r^{l(r)}.$$

If for some prime factor p of n the number $p - 1$ is B-smooth, then $p - 1$ divides Q (observe that $r^{l(r)}$ is the largest power of r that is $\leq n$). Since $a^{p-1} \equiv 1 \pmod{p}$, we obtain $a^Q \equiv 1 \pmod{p}$, and so

$$e = \gcd(b - 1, n) = \gcd(a^Q - 1, n) \geq p$$

in Step 4. Thus, Step 5 is successful unless we are very unlucky and $e = n$.

A good value of B has to be selected by a careful trade-off. If B is small, then the computations in the algorithm are faster, but the probability of success will be low. If B is not too small, then the likelihood of $p-1$ being B-smooth is quite large. Practitioners say that with experience and *feng shui* one gets the knack for the proper choice of B.

Finally, we describe the *Pollard rho algorithm.* Let S be a finite set with $m \geq 2$ elements, let $f : S \to S$ be a self-map of S, and let $s_0, s_1, \dots$ be a sequence of elements of S defined recursively by $s_i = f(s_{i-1})$ for $i = 1, 2, \dots$ with an arbitrary initial value s_0. Since S is finite, this sequence is ultimately periodic. In particular, there exist subscripts i and j with $0 \leq i < j$ such that $s_i = s_j$, and so there are repeated terms.

Lemma 2.3.19 *For a real number $\lambda > 0$ put $\ell = 1 + \lfloor \sqrt{2\lambda m} \rfloor$. Then the fraction of ordered pairs (f, s_0) such that all elements $s_0, s_1, \dots, s_\ell$ are different is smaller than $e^{-\lambda}$.*

Proof. We may assume that $\ell < m$. The total number of all ordered pairs (f, s_0) is m^{m+1} and the number of ordered pairs (f, s_0) with different $s_0, s_1, \dots, s_\ell$ is $m^{m-\ell} \prod_{j=0}^{\ell} (m-j)$. Hence the fraction in question is

$$h(m, \ell) := m^{-\ell-1} \prod_{j=0}^{\ell} (m-j) = \prod_{j=0}^{\ell} \left(1 - \frac{j}{m}\right).$$

Because of $\log(1-u) < -u$ for $0 < u < 1$, we get

$$\log h(m, \ell) = \sum_{j=0}^{\ell} \log\left(1 - \frac{j}{m}\right) < -\frac{1}{m} \sum_{j=0}^{\ell} j = -\frac{\ell(\ell+1)}{2m} < -\frac{\ell^2}{2m} < -\lambda$$

and thus $h(m, \ell) < e^{-\lambda}$. □

Lemma 2.3.20 *The expected value of the smallest integer $\ell \geq 1$ with $s_\ell = s_{2\ell}$ has an order of magnitude at most $m^{1/2}$.*

Proof. Denote by ℓ_1 and ℓ_2 the length of the preperiod and the length of the period of the sequence $s_0, s_1, \dots$, respectively. Then for $\ell = \ell_2(1 + \lfloor \ell_1/\ell_2 \rfloor) > \ell_1$ we obtain $s_\ell = s_{2\ell}$. By Lemma 2.3.19, the expected value of $\ell \leq \ell_1 + \ell_2$ has an order of magnitude at most $m^{1/2}$. □

Now let p be a prime factor of a given composite integer n. The Pollard rho algorithm tries to find repeated terms in the sequence $a_0, a_1, \dots$ of elements of $S =$

$Z_p = \{0, 1, \dots, p-1\}$ generated by $a_0 = 2$ and $a_i = f(a_{i-1})$ for $i = 1, 2, \dots$, where $f(s) \in Z_p$ is determined by $f(s) \equiv s^2 + 1 \pmod{p}$ for all $s \in Z_p$. Since p divides n but is unknown, this is effected by carrying out the analogous computation in Z_n and testing whether $\gcd(a_i - a_{2i}, n) > 1$. If also $\gcd(a_i - a_{2i}, n) < n$, then a nontrivial factor of n has been found. In the implementation, the terms a_i and $b_i := a_{2i}$ are computed in parallel.

Algorithm 2.3.21 (Pollard Rho Algorithm) Find a nontrivial factor of the composite integer n.
Step 1: Put $a_0 = 2$, $b_0 = 2$.
Step 2: For $i = 1, 2, \dots$ do the following:
(a) Compute $a_i \equiv a_{i-1}^2 + 1 \pmod{n}$ and $b_i \equiv (b_{i-1}^2 + 1)^2 + 1 \pmod{n}$ with $a_i, b_i \in Z_n$.
(b) Compute $d_i = \gcd(a_i - b_i, n)$.
(c) If $1 < d_i < n$, then d_i is a nontrivial factor of n and stop with success. If always $d_i = 1$ or n for i up to a prescribed bound, then stop with failure.

Remark 2.3.22 Under the assumption that $x^2 + 1$ behaves like a random function modulo p, we can apply Lemma 2.3.20 with $m = p$. Since there always exists a prime factor p of n with $p \le n^{1/2}$, we can therefore expect that Algorithm 2.3.21 terminates with success after $O(n^{1/4})$ steps. Here and later, $g(n) = O(h(n))$ is equivalent to the existence of a positive constant C such that $|g(n)| \le Ch(n)$ for all positive integers n, where g is a real-valued function and h is a nonnegative function on $\mathbb{N}$. In the rare case where the algorithm fails, we replace $x^2 + 1$ by a function $x^2 + c$ with a new value for the constant c, such as $c = 2$ or $c = 3$.

Example 2.3.23 Let us factor $n = pq = 1927$ by the Pollard rho algorithm. We summarize the computation in the following table.

i	0	1	2	3	4	5	6	7
a_i	2	5	26	677	1631	902	411	1273
b_i	2	26	1631	411	1850	1005	205	535
d_i	$-$	1	1	1	1	1	1	41

Therefore n has the prime factor 41, and by division we obtain $n = 41 \cdot 47$.

Can you guess where the name "rho algorithm" comes from? This is not exactly a million dollar question, but still the answer is not obvious. Let us return to Example 2.3.23 and compute the a_i modulo the prime factor 47. This initially yields the terms $2, 5, 26, 19, 33, 9, 35, 4, 17, 8, 18, 43, 17$. From then on the sequence cycles since $a_{12} \equiv a_8 \equiv 17 \pmod{47}$. If you picture the situation starting for example from 19 (see Figure 2.5), then you see the Greek letter rho appearing! No kidding, this is the reason for the name of the algorithm.

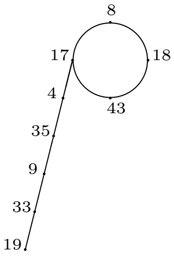

Figure 2.5: The Pollard rho algorithm

2.4 Cryptosystems based on discrete logarithms

2.4.1 The cryptosystems

The factorization problem for large integers is not the only number-theoretic problem that serves as the basis for a public-key cryptosystem. The discrete logarithm problem also gets top billing in the area. Let us introduce the discrete logarithm function for finite fields without further ado. A crucial role is played by the concept of a primitive element of a finite field (see Definition 1.4.34).

Definition 2.4.1 Let $\mathbb{F}_q$ be a finite field and let $g \in \mathbb{F}_q^*$ be a primitive element of $\mathbb{F}_q$. For each $a \in \mathbb{F}_q^*$, the unique integer h with $0 \leq h \leq q-2$ such that $g^h = a$ is called the *discrete logarithm* (or the *index*) $\mathrm{ind}_g(a)$ of a to the base g.

Example 2.4.2 Note that $2 \in \mathbb{F}_5^*$ is a primitive element of $\mathbb{F}_5$, and in this case the discrete logarithm has the following values:

$$\mathrm{ind}_2(1) = 0,\ \mathrm{ind}_2(2) = 1,\ \mathrm{ind}_2(3) = 3,\ \mathrm{ind}_2(4) = 2.$$

It would of course be more natural to denote the discrete logarithm function by “log”, but since this notation is already reserved for the logarithm function in calculus, we use “ind” instead. The *discrete logarithm problem* is the problem of computing $\mathrm{ind}_g(a)$ for a finite field $\mathbb{F}_q$ of large order q. There are easily obtained values of the discrete logarithm like $\mathrm{ind}_g(1) = 0$ and $\mathrm{ind}_g(g) = 1$, but in general the discrete logarithm problem is believed to be difficult. In fact, practical experience shows that the discrete logarithm problem for $\mathbb{F}_q$ is about as hard as factoring an integer that has roughly the same size as q.

The discrete logarithm for $\mathbb{F}_q$ has similar properties as the ordinary logarithm, but identities have to be replaced by congruences modulo $q-1$. In detail, if $a, b \in \mathbb{F}_q^*$ and $n \in \mathbb{N}$, then

$$\mathrm{ind}_g(ab) \equiv \mathrm{ind}_g(a) + \mathrm{ind}_g(b) \pmod{q-1},$$
$$\mathrm{ind}_g(a^n) \equiv n\ \mathrm{ind}_g(a) \pmod{q-1}.$$

If the discrete logarithm problem for $\mathbb{F}_q$ is presumed to be difficult, then we can view the discrete exponential function $h \in Z_{q-1} \mapsto g^h \in \mathbb{F}_q^*$ as a one-way function (compare with Definition 2.3.2). Note that the computation of g^h, even for very large exponents h, can be carried out quickly by Algorithm 2.3.9. Therefore the discrete exponential function can serve as a basis for public-key cryptosystems.

The first cryptographic scheme of this type that we discuss (and also historically the first) is actually not a public-key cryptosystem for encryption, but a key-exchange (or key-agreement) scheme which can be part of a hybrid cryptosystem (see Subsection 2.3.1). This scheme was introduced in the seminal paper of Diffie and Hellman [40]. The objective is to exchange (or agree on) a cryptographic key, for instance for a symmetric block cipher, over an insecure channel. All participants of a communication system share a large finite field $\mathbb{F}_q$ and a primitive element g of $\mathbb{F}_q$. Let us describe how the participants Alice and Bob establish a common key.

Algorithm 2.4.3 (Diffie-Hellman Key Exchange) Alice and Bob want to establish a common key, given the large finite field $\mathbb{F}_q$ and the primitive element g of $\mathbb{F}_q$.
Step 1: Alice chooses a random integer h with $2 \leq h \leq q-2$ and Bob chooses a random integer k with $2 \leq k \leq q-2$.
Step 2: Alice sends $g^h \in \mathbb{F}_q$ to Bob over the channel, while Bob sends $g^k \in \mathbb{F}_q$ to Alice over the channel.
Step 3: The common key is $g^{hk} \in \mathbb{F}_q$, which Alice computes as $(g^k)^h$ and Bob computes as $(g^h)^k$.

An observer of the communication sees g^h and g^k going over the channel. If he is a malicious adversary, then he will try to figure out g^{hk}, given g^h and g^k. But according to the current know-how, the only way to do that is to first compute h and k, and so he has to solve the discrete logarithm problem for $\mathbb{F}_q$. If q is large (say q has at least 300 decimal digits), then the Diffie-Hellman key-exchange scheme is considered secure.

There is also a public-key cryptosystem based on the presumed difficulty of the discrete logarithm problem, but it was designed almost a decade after the Diffie-Hellman paper [40] since it is not so obvious here how to build trapdoor information into the system. Actually, a new idea was needed, namely to use a random quantity in the encryption algorithm. Just like the work of Diffie, this cryptosystem is also the achievement of a graduate student at Stanford University.

Our typical user Bob chooses a large finite field $\mathbb{F}_q$ and a primitive element g of $\mathbb{F}_q$. Then he selects an integer h with $2 \leq h \leq q-2$ and efficiently computes $a = g^h \in \mathbb{F}_q^*$ by Algorithm 2.3.9.

Algorithm 2.4.4 (ElGamal Cryptosystem) The public key of Bob is the ordered triple (q, g, a) and the private key of Bob is $h = \mathrm{ind}_g(a)$. The plaintext source is $\mathbb{F}_q^*$.
Encryption: Suppose that Alice wants to send a plaintext $m \in \mathbb{F}_q^*$ to Bob. Then

Alice chooses a random integer r with $2 \leq r \leq q-2$ and computes $c_1 = g^r \in \mathbb{F}_q^*$ and $c_2 = ma^r \in \mathbb{F}_q^*$. The ordered pair (c_1, c_2) is sent as the ciphertext to Bob.
Decryption: Upon receiving the ciphertext (c_1, c_2), Bob computes $c_2(c_1^h)^{-1} = m$ and thus recovers the plaintext.

The security analysis for the ElGamal cryptosystem is similar to that for the Diffie-Hellman key-exchange scheme. The result is that this public-key cryptosystem is considered secure if q has at least 300 decimal digits.

Remark 2.4.5 In practical implementations of the Diffie-Hellman key-exchange scheme and the ElGamal cryptosystem, one usually takes for $\mathbb{F}_q$ a finite prime field $\mathbb{F}_p$ with a prime number p (see Theorem 1.4.5 and Remark 1.4.6). The primitive element g then becomes a primitive root modulo p. However, for the theory it does not make any difference whether we use a general finite field or a finite prime field. This brings to mind a famous saying by the legendary baseball coach Yogi Berra: "In theory there is no difference between theory and practice, in practice there is."

2.4.2 Computing discrete logarithms

For special prime powers q or prime powers q that are not too large, discrete logarithms for $\mathbb{F}_q$ can be computed. We present three algorithms of this type.

A first rather simple algorithm for computing discrete logarithms for $\mathbb{F}_q$ uses about $q^{1/2}$ operations in $\mathbb{F}_q$. It is known by the colorful name *baby-step giant-step algorithm*. For all $c \in \mathbb{F}_q^*$ and all integers $n \geq 0$, we use the notation $c^{-n} := (c^{-1})^n$.

Algorithm 2.4.6 (Baby-Step Giant-Step Algorithm) Let $a \in \mathbb{F}_q^*$ and let g be a primitive element of the finite field $\mathbb{F}_q$. Compute $\mathrm{ind}_g(a)$.
Step 1: Put $m := \lceil \sqrt{q-1} \rceil$ and set up the table (this is the baby step)

j	0	1	2	$\ldots$	$m-1$
g^j	g^0	g^1	g^2	$\ldots$	g^{m-1}

Step 2: Compute g^{-m} and put $a_0 := a$.
Step 3: For $i = 0, 1. \ldots, m-1$ do the following:
(a) Check whether a_i occurs in the second row of the above table and read the corresponding j above it.
(b) If yes, put $\mathrm{ind}_g(a) = im + j$ and stop. Otherwise put $a_{i+1} := a_i g^{-m}$ (this is the giant step) and return to (a).

It remains to show that the algorithm really calculates the discrete logarithm $\mathrm{ind}_g(a)$. By division with remainder, we can write $\mathrm{ind}_g(a) = im + j$ with integers $0 \leq i \leq m-1$ and $0 \leq j \leq m-1$. Then from $g^{im+j} = a$ we obtain

$$g^j = ag^{-im} = a_0 g^{-im} = a_1 g^{-(i-1)m} = \cdots = a_{i-1} g^{-m} = a_i,$$

and so in Step 3(a) the element a_i corresponds to j.

Since a discrete logarithm $h = \mathrm{ind}_g(a)$ for $\mathbb{F}_q$ satisfies $0 \leq h \leq q-2$, it suffices to compute h modulo $q-1$. The case $q = 2$ is trivial, and so we can assume that $q \geq 3$. By the Chinese remainder theorem (see Theorem 1.2.9), we can proceed by determining h modulo all prime powers in the canonical factorization of $q-1$. This is the strategy of the *Silver-Pohlig-Hellman algorithm*. Let

$$q - 1 = \prod_{j=1}^{k} p_j^{e_j}$$

be the canonical factorization of $q-1$ and let $a \in \mathbb{F}_q^*$ and a primitive element g of $\mathbb{F}_q$ be given.

We take a typical prime factor $p = p_j$ of $q-1$. The idea is to first compute the least residue of $h = \mathrm{ind}_g(a)$ modulo p, which is h_0, say. Then $h = h_0 + kp$ for some $k \in \mathbb{Z}$, and so

$$\frac{q-1}{p} h \equiv \frac{q-1}{p}(h_0 + kp) \equiv \frac{q-1}{p} h_0 \pmod{q-1}.$$

Therefore

$$g^{(q-1)h_0/p} = g^{(q-1)h/p} = (g^h)^{(q-1)/p} = a^{(q-1)/p}.$$

Thus, we compute $b := g^{(q-1)/p}$ and then $b^0, b^1, b^2, \ldots$ until we get $a^{(q-1)/p}$. This will happen at b^{h_0}, where $0 \leq h_0 \leq p-1$, and so h_0 is determined.

If there are higher powers of p dividing $q-1$, say p^2 divides $q-1$, then we determine the least residue of h modulo p^2, which has the form $h_0 + h_1 p$ with some $h_1 \in Z_p$. To do this, we put $a_1 = ag^{-h_0}$ and compute the exponent $h_1 \in Z_p$ such that $b^{h_1} = a_1^{(q-1)/p^2}$. For even higher powers of p dividing $q-1$, we continue this procedure in a similar way. Finally, h is determined modulo all prime powers in the canonical factorization of $q-1$, and so h is uniquely determined by the Chinese remainder theorem.

In each of the above steps we have to compute at worst $b^0, b^1, \ldots, b^{p-1}$, so if all prime factors p of $q-1$ are relatively small, then the Silver-Pohlig-Hellman algorithm is feasible.

Example 2.4.7 Here is a toy example for the Silver-Pohlig-Hellman algorithm with $q = 19$, $g = 2$, and $a = 6$. Note that $q - 1 = 18 = 2 \cdot 3^2$. We first consider the more involved case of the prime factor $p = 3$. We determine $b = g^{(q-1)/p}$ from $b \equiv 2^6 \equiv 64 \equiv 7 \pmod{19}$, and so $b = 7 \in \mathbb{F}_{19}$. Next we compute $a^{(q-1)/p}$ by

$$a^{(q-1)/p} \equiv 6^6 \equiv (6^2)^3 \equiv (-2)^3 \equiv -8 \equiv 11 \pmod{19},$$

and so $a^{(q-1)/p} = 11 \in \mathbb{F}_{19}$. Now we calculate $b^0, b^1, b^2, \ldots$ in $\mathbb{F}_{19}$ until we obtain $a^{(q-1)/p}$. We get $b^0 = 1$, $b^1 = 7$, $b^2 = 11$, and so $h_0 = 2$. Since 3^2 divides $q-1$, another step is needed. If $h = \mathrm{ind}_g(a)$, then

$$h \equiv h_0 + h_1 p \equiv 2 + 3h_1 \pmod{9}.$$

Note that $a_1 = ag^{-h_0} = 11 \in \mathbb{F}_{19}$ and that $h_1 \in Z_3$ is the exponent such that $b^{h_1} = a_1^{(q-1)/p^2} = 11^2 = 7 \in \mathbb{F}_{19}$. Hence $h_1 = 1$ since $b = 7$. Therefore $h \equiv 2 + 3 \cdot 1 \equiv 5 \pmod 9$. The case of the prime factor $p = 2$ is easy since now

$$a^{(q-1)/p} \equiv 6^9 \equiv (6^2)^4 \cdot 6 \equiv (-2)^4 \cdot 6 \equiv 1 \pmod{19},$$

and so $h \equiv 0 \pmod 2$. Since $0 \leq h \leq 17$, the congruence $h \equiv 5 \pmod 9$ implies that $h = 5$ or 14, and so $h \equiv 0 \pmod 2$ yields $h = \mathrm{ind}_g(a) = \mathrm{ind}_2(6) = 14$.

Finally, we present a somewhat more elaborate algorithm for computing discrete logarithms, the *index-calculus algorithm*. We restrict the discussion to the case of a finite prime field $\mathbb{F}_p$, but see Remark 2.4.9 below. Let p be a prime number and let g be a primitive root modulo p (or equivalently a primitive element of $\mathbb{F}_p$). Let B be an integer with $2 \leq B < p$ and recall the notion of a B-smooth integer from Subsection 2.3.3.

In the first step of the index-calculus algorithm, we determine $\mathrm{ind}_g(r)$ for all prime numbers $r \leq B$. To do this, we choose a random integer m with $1 \leq m \leq p-2$ and compute the least residue of g^m modulo p. If this least residue is B-smooth, then

$$g^m \equiv \prod_{r \leq B} r^{e(r)} \pmod p.$$

Otherwise, we pick a new m. By the rules for discrete logarithms, we obtain

$$m \equiv \sum_{r \leq B} e(r)\, \mathrm{ind}_g(r) \pmod{p-1}.$$

This is a linear congruence for the unknowns $\mathrm{ind}_g(r)$. By producing sufficiently many of these congruences, we hope that the resulting system will have a unique solution modulo $p-1$.

In the second step of the index-calculus algorithm, let a be an integer with $\gcd(a, p) = 1$ for which we want to calculate $\mathrm{ind}_g(a)$. (Strictly speaking, we are really talking about the least residue of a modulo p, viewed as an element of $\mathbb{F}_p^*$, but this does not make any difference in the computations.) We choose a random integer s with $0 \leq s \leq p-2$ and compute the least residue of ag^s modulo p. If this least residue is B-smooth, then

$$ag^s \equiv \prod_{r \leq B} r^{f(r)} \pmod p.$$

Otherwise, we pick a new s. By the rules for discrete logarithms, we get

$$\mathrm{ind}_g(a) + s \equiv \sum_{r \leq B} f(r)\, \mathrm{ind}_g(r) \pmod{p-1}.$$

Since all discrete logarithms $\mathrm{ind}_g(r)$ for prime numbers $r \leq B$ have been computed in the first step, this determines $\mathrm{ind}_g(a)$ uniquely.

The index-calculus algorithm is feasible if p is not too large, since then there is a higher chance that the least residues modulo p we have to calculate are B-smooth.

Example 2.4.8 We compute $\mathrm{ind}_2(6)$ in Example 2.4.7 by the index-calculus algorithm. Since $a = 6 = 2 \cdot 3$, it suffices to take $B = 3$. In the first step, we have to compute $\mathrm{ind}_2(2)$ and $\mathrm{ind}_2(3)$. Clearly $\mathrm{ind}_2(2) = 1$. For the determination of $\mathrm{ind}_2(3)$, we are obliged to find a suitable power $g^m = 2^m$ for which the least residue modulo 19 is 3-smooth. Now the least residues of 2, 2^2, 2^3, and 2^4 modulo 19 are 3-smooth, but they do not involve the number 3. Next $2^5 \equiv 13 \pmod{19}$, $2^6 \equiv 7 \pmod{19}$, and $2^7 \equiv 14 \pmod{19}$ yield least residues modulo 19 that are not 3-smooth. But $2^8 \equiv 9 \pmod{19}$ has a 3-smooth least residue modulo 19. Taking the discrete logarithm to the base 2 in $2^8 = 3^2$ in $\mathbb{F}_{19}$, we obtain $8 \equiv 2\,\mathrm{ind}_2(3) \pmod{18}$, and so $\mathrm{ind}_2(3) \equiv 4 \pmod{9}$. The value $\mathrm{ind}_2(3) = 4$ is not possible since $2^4 \equiv 16 \not\equiv 3 \pmod{19}$, and so necessarily $\mathrm{ind}_2(3) = 13$. The second step of the index-calculus algorithm is easy in this case. We simply choose $s = 0$, then $ag^s \equiv 6 \equiv 2 \cdot 3 \pmod{19}$, and taking the discrete logarithm to the base 2 we obtain

$$\mathrm{ind}_g(a) \equiv \mathrm{ind}_2(2 \cdot 3) \equiv \mathrm{ind}_2(2) + \mathrm{ind}_2(3) \equiv 14 \pmod{18}.$$

Thus, the final answer is $\mathrm{ind}_g(a) = \mathrm{ind}_2(6) = 14$ as in Example 2.4.7.

Remark 2.4.9 For a finite field $\mathbb{F}_q$ where q is not a prime number, say $q = p^k$ with a prime number p and an integer $k \geq 2$, there is also an index-calculus algorithm. The computations are then carried out not in $\mathbb{Z}$, but in the polynomial ring $\mathbb{F}_p[x]$. Furthermore, $\mathbb{F}_q$ is identified with the residue class field $\mathbb{F}_p[x]/(v(x))$, where $v(x) \in \mathbb{F}_p[x]$ is irreducible over $\mathbb{F}_p$ with $\deg(v(x)) = k$ (compare with Subsection 1.4.3). We refer to [103, Section 9.3] for details of the algorithm.

2.5 Digital signatures

2.5.1 Digital signatures from public-key cryptosystems

Let us consider confidential communication between two users A and B with encryption by a public-key cryptosystem. Since the encryption key of B is public, anybody can send an encrypted message to B. How can B be sure that the message really came from A? This question is very important in highly sensitive areas such as legal and financial matters. The recipient (for example a bank) has to be absolutely sure that the message (asking for example for a money withdrawal) comes from an authorized person (for example the owner of the bank account). Conventionally, one uses handwritten signatures for this purpose. In electronic communications one employs digital signatures. Thus, digital signatures are essential in e-banking, e-commerce, e-government, and anything else starting with “e-”.

Digital signatures are provided by signature schemes. A *signature scheme* consists of two algorithms:
(i) a *signing algorithm* sig_K, depending on a signature key K, which computes for each possible message m a corresponding signature $s = \mathrm{sig}_K(m)$ that is appended to the message;
(ii) a *verification algorithm* $\mathrm{ver}_{\overline{K}}$ which checks, for all possible messages m and all possible signatures s, whether $s = \mathrm{sig}_K(m)$. Thus,

$$\mathrm{ver}_{\overline{K}}(m, s) = \begin{cases} \text{true} & \text{if } s = \mathrm{sig}_K(m), \\ \text{false} & \text{if } s \neq \mathrm{sig}_K(m). \end{cases}$$

The signing and verification algorithms should be fast. The signature key K and so the function sig_K are secret, whereas $\mathrm{ver}_{\overline{K}}$ is a public function, with a public verification key $\overline{K}$, so that anybody can check digital signatures. It should be computationally infeasible to forge a digital signature on a message m. That is, given m, only the authorized user A should be able to compute the signature s such that $\mathrm{ver}_{\overline{K}}(m, s) = \text{true}$.

It is worth emphasizing that a handwritten signature is independent of the document to be signed, whereas a digital signature depends on the message m. The reason for the latter is that the physical link between document and handwritten signature has to be replaced by a logical link between message and digital signature.

Certain public-key cryptosystems can be used to produce digital signatures. The public-key cryptosystem must satisfy the following condition, in addition to PKC1, PKC2, and PKC3 in Subsection 2.3.1.

PKC4: The decryption algorithm D can be applied to every plaintext M and $E(D(M)) = M$ for all M.

Note that normally the decryption algorithm is applied only to ciphertexts, so PKC4 is an additional property that needs to be checked. An example of a public-key cryptosystem that does not satisfy PKC4 is the ElGamal cryptosystem in Algorithm 2.4.4, since there the plaintexts are elements of $\mathbb{F}_q^*$ and the ciphertexts are ordered pairs of elements of $\mathbb{F}_q^*$.

If a public-key cryptosystem satisfies PKC4, then we get a signature scheme as follows. We consider again our two acquaintances Alice and Bob. Suppose that Bob wants to sign the message M that he sends to Alice. Then the following two steps are executed, with the notation in Subsection 2.3.1.
(i) *Signing algorithm*: Bob takes his secret decryption key K' and computes the message-dependent signature $S = D_{K'}(M)$.
(ii) *Verification algorithm*: Alice looks up the public encryption key K of Bob. Then she takes the signature S and computes $E_K(S) = E_K(D_{K'}(M))$. If the result is M, then the signature is verified, otherwise it is rejected. Alice can be satisfied that the message M came from Bob since no other person would have used the secret key K'

of Bob to compute $S = D_{K'}(M)$.

An important example is provided by the RSA cryptosystem. Remember that we have to check the property PKC4 above. In the RSA cryptosystem, decryption is achieved by the map that sends $c \in Z_n$ to the least residue of c^d modulo n. Obviously, this decryption algorithm can be applied to every plaintext $m \in Z_n$. Furthermore, the corresponding encryption algorithm computes $c \equiv m^e \pmod{n}$ with $ed \equiv 1 \pmod{\phi(n)}$. Then

$$E(D(m)) \equiv E(m^d) \equiv (m^d)^e \equiv m^{ed} \equiv m \pmod{n}$$

for all $m \in Z_n$, where the last step follows by Lemma 2.3.6. Thus, the RSA cryptosystem satisfies the property PKC4 and can be used for digital signatures.

In order to set up the RSA signature scheme, the prime numbers p and q and the integers $n = pq$, e, and d are chosen as in the RSA cryptosystem, and the public and private keys are now those of Bob in the role of the signer.

Algorithm 2.5.1 (RSA Signature Scheme) The public key of Bob is the ordered pair (n, e) and the private key of Bob is the ordered triple (p, q, d).
Signing: Bob signs a plaintext $m \in Z_n$ by computing $s \in Z_n$ with $s \equiv m^d \pmod{n}$ and sends the ordered pair (m, s) to Alice.
Verification: Upon receiving (m, s), Alice computes the least residue of s^e modulo n and checks whether it agrees with m.

Clearly, the security level of the RSA signature scheme is exactly the same as for the RSA cryptosystem. Note that we have just turned the tables: encryption has become verification and decryption has become signing.

The ElGamal cryptosystem cannot be used directly for digital signatures since we have already noted that it does not satisfy PKC4, but there is a slight modification that works. In order to set up the *ElGamal signature scheme*, the signer Bob chooses a large prime number p, a primitive root g modulo p, and an integer h with $2 \le h \le p - 2$. Then Bob computes $a \in Z_p$ with $a \equiv g^h \pmod{p}$.

Algorithm 2.5.2 (ElGamal Signature Scheme) The public key of Bob is the ordered triple (p, g, a) and the private key of Bob is h.
Signing: Bob signs a plaintext $m \in Z_p$ by choosing a random integer r with $2 \le r \le p - 2$ and $\gcd(r, p - 1) = 1$, computing $b \in Z_p$ with $b \equiv g^r \pmod{p}$, and sending the ordered triple (m, b, c) to Alice, where $c \in Z_{p-1}$ is the unique solution of the congruence

$$rc \equiv m - bh \pmod{p - 1}.$$

Verification: Upon receiving (m, b, c), Alice computes the least residue of $a^b b^c$ modulo p and checks whether it agrees with the least residue of g^m modulo p.

It remains to prove that $a^b b^c \equiv g^m \pmod{p}$. This holds since

$$a^b b^c \equiv g^{hb} g^{rc} \equiv g^{hb} g^{m-bh} \equiv g^m \pmod{p}.$$

As for the ElGamal cryptosystem, the security of the ElGamal signature scheme is based on the presumed difficulty of the discrete logarithm problem, in this case for the finite prime field $\mathbb{F}_p$.

In practice, one wants to combine signing and public-key encryption. The crucial question is about the proper order of these operations: do you first sign or first encrypt? The answer becomes obvious once you pose the question in the analog world: when you mail for example a confidential contract like a job contract, would you first sign the contract and then put it into an envelope that you seal, or would you put the unsigned contract into an envelope and then sign the sealed envelope? Consequently, you first sign and then you encrypt the combined plaintext and signature. The legitimate receiver Alice undoes these operations in the correct order: she first decrypts and then she verifies the signature attached to the plaintext. If you carry out these steps in the wrong order, that is, if you first encrypt and then sign, then an adversary having access to the insecure channel can replace your signature by his own signature and send the ciphertext-signature pair to Alice. When Alice applies the verification algorithm of the adversary, everything checks and Alice will conclude that the message originated with the adversary.

2.5.2 DSS and related schemes

The *Digital Signature Standard* (*DSS*) is a signature scheme that was adopted as a standard by the U.S. Government in 1994, thus sanctioning a design by one of its own agencies. The DSS is a modification of the ElGamal signature scheme in Algorithm 2.5.2.

Note that the signature in the ElGamal signature scheme is an ordered pair (b, c), where b and c are integers modulo p and $p-1$, respectively. In 1994 it was already necessary to choose p as a 512-bit prime number in order to make the ElGamal signature scheme secure. Thus, an ElGamal signature can be expected to have up to 1024 bits. This is too long for typical applications such as smart cards. Nowadays it would be preferable to choose a 1024-bit prime modulus p, leading to even longer 2048-bit ElGamal signatures in the worst case. For this reason, the ElGamal signature scheme was not used directly.

The DSS signs 160-bit messages with 320-bit signatures, but the computations are performed with a prime modulus p that has between 512 and 1024 bits. This is achieved by replacing the primitive root g modulo p in the ElGamal signature scheme by a nonzero integer $g_1 \in Z_p$ such that the multiplicative order of g_1 modulo p is equal to p_1, where p_1 is a 160-bit prime number dividing $p-1$. If a primitive root g modulo p is known, then such an integer g_1 can be obtained by the congruence

$g_1 \equiv g^{(p-1)/p_1} \pmod{p}$. The signer Bob chooses an integer h with $2 \leq h \leq p_1 - 2$ and computes $a \in Z_p$ with $a \equiv g_1^h \pmod{p}$.

Algorithm 2.5.3 (Digital Signature Standard) The public key of Bob is the ordered quadruple (p, p_1, g_1, a) and the private key of Bob is h.
Signing: Bob signs a 160-bit plaintext m by choosing a random integer r with $2 \leq r \leq p_1 - 2$, computing the least residue of g_1^r modulo p, and then computing the least residue b of that number modulo p_1. Next Bob determines the unique solution $c \in Z_{p_1}$ of the congruence

$$rc \equiv m + bh \pmod{p_1}.$$

In the rare case where $c = 0$, a new random integer r is chosen. Finally Bob sends the ordered triple (m, b, c) to Alice.
Verification: Upon receiving (m, b, c), Alice first finds the solutions $e_1, e_2 \in Z_{p_1}$ of the congruences $ce_1 \equiv m \pmod{p_1}$ and $ce_2 \equiv b \pmod{p_1}$. Then she computes the least residue of $g_1^{e_1} a^{e_2}$ modulo p, then the least residue of that number modulo p_1, and finally she checks whether the latter number agrees with b.

It remains to prove that $g_1^{e_1} a^{e_2} \equiv g_1^r \pmod{p}$. Note that

$$g_1^{e_1} a^{e_2} \equiv g_1^{e_1} g_1^{he_2} \equiv g_1^{e_1 + he_2} \pmod{p}.$$

Furthermore,

$$c(e_1 + he_2) \equiv m + bh \equiv cr \pmod{p_1}.$$

From $\gcd(c, p_1) = 1$ we obtain $e_1 + he_2 \equiv r \pmod{p_1}$, and since the multiplicative order of g_1 modulo p is p_1, we get indeed $g_1^{e_1} a^{e_2} \equiv g_1^r \pmod{p}$.

Remark 2.5.4 We can sign only 160-bit messages with the DSS, but in practice messages can be megabytes in size. This raises the question of how to sign long messages with the DSS. Of course, we could split up a long message into 160-bit chunks and then sign each chunk separately. But this has several disadvantages: (i) the resulting signature is enormous, namely about twice as long as the message; (ii) the communication is slowed down by the time it takes to compute many signatures; (iii) a loss of security is possible since an adversary could rearrange or remove various chunks of a signed message and the resulting message plus signature would still be verified. The preferred solution for the signing of long messages with the DSS is to "hash" the message to a 160-bit message digest and then to sign this message digest with the DSS. In the communication, the original message is sent together with the signed message digest. For the purpose of "hashing", we need a *hash function*, that is, a function mapping long strings of symbols into much shorter strings of symbols. In the present case, the hash function must be publicly known so that everyone can compute the message digest in order to verify the signature. For security reasons, hash functions with special cryptographic properties have to be used. We refer to [116, Chapter 9] and [160, Chapter 6] for information on hash functions.

It seems that the DSS is unnecessarily complicated (maybe this is typical of something produced by a huge bureaucracy), and so several simpler digital signature schemes using the same basic idea were proposed. We describe a particularly elegant alternative, the *Nyberg-Rueppel signature scheme*. The setup for the Nyberg-Rueppel signature scheme is the same as for the DSS, except that there are no constraints on the sizes of p and p_1.

Algorithm 2.5.5 (Nyberg-Rueppel Signature Scheme) The public key of Bob is the ordered quadruple (p, p_1, g_1, a) and the private key of Bob is h.
Signing: Bob signs a plaintext $m \in \mathbb{F}_p$ by choosing a random integer r with $2 \leq r \leq p_1 - 2$, computing $b = mg_1^{-r} \in \mathbb{F}_p$, and sending the ordered triple (m, b, c) to Alice, where $c = bh + r \in \mathbb{F}_{p_1}$.
Verification: Upon receiving (m, b, c), Alice computes $bg_1^c a^{-b} \in \mathbb{F}_p$ and checks whether it agrees with m.

It remains to prove that $bg_1^c a^{-b} = m$ in $\mathbb{F}_p$. This holds since

$$bg_1^c a^{-b} = bg_1^{bh+r} g_1^{-bh} = bg_1^r = m.$$

2.6 Threshold schemes

The need for threshold schemes is best explained by an example. The following is a standard operating principle in banks (the “four-eyes principle”): in order to open the bank’s vault, at least two senior employees have to cooperate; one person is not enough. Thus, we require a scheme to distribute the vault’s lock combination such that two authorized persons can generate the lock combination, but one person cannot. The four-eyes principle is used also in other sensitive areas, such as the control of nuclear weapons.

In a more general and abstract setting, the relevant cryptographic scheme is described as follows. Let n be the number of users of the scheme. Let S be the secret that needs to be protected, for example the key for a cryptosystem or a lock combination. The n users receive data $S_1, \ldots, S_n$, respectively, which may be thought of as partial information about S and are called the “shares”. The idea is that certain coalitions of users can reconstruct the secret S from their shares. Such a general scheme is called a *secret-sharing scheme*. The shares are customarily generated by a trusted authority which also distributes them to the users. A threshold scheme is a special type of secret-sharing scheme.

Definition 2.6.1 Let k and n be integers with $2 \leq k \leq n$. A secret-sharing scheme with threshold k and n users is called a (k, n)-*threshold scheme* if it has the following properties:
(i) any k or more users can reconstruct the secret from their shares;
(ii) for $k - 1$ or fewer users it is impossible to reconstruct the secret.

Remark 2.6.2 In the above example from banking, the threshold is $k = 2$ and n is the number of employees authorized to open the vault. Thus, we need a $(2, n)$-threshold scheme to implement the four-eyes principle.

Remark 2.6.3 Here is a simplistic threshold scheme, say for $k = 2$ and $n = 2$. Let us be concrete and assume that the lock combination of the vault consists of six decimal digits. We give the first three digits to the president of the bank and the other three digits to the vice-president. On first glance, it looks as if the definition of a $(2, 2)$-threshold scheme were satisfied. However, if say the vice-president is dishonest, then he has to guess only the first three digits of the lock combination rather than the full six digits. It is feasible for him to try the 10^3 possibilities during a long night session or over a weekend, but it is impossible for him to try 10^6 possibilities unless he is unbelievably lucky and beats the chance of 10^{-6} within a manageable time span. Hence this simplistic scheme does not work since it dramatically reduces the time that is needed for an exhaustive search. In a well-designed threshold scheme, the shares should contain about as much unknown information or uncertainty as the secret.

Example 2.6.4 Threshold schemes can be used also to protect against loss of information. If S is a piece of information that we want to protect, then we can use a (k, n)-threshold scheme to distribute partial information about S to n users. Even if $n - k$ of these shares are lost or destroyed, we can still recover S from the remaining k shares. The threshold k also allows up to $k - 1$ of the shares to be disclosed by breaches of security, without compromising S. This has obvious applications in highly exposed or dangerous environments such as a battle field.

We describe a number-theoretic (k, n)-threshold scheme, the *Shamir threshold scheme*, which is named after its designer Adi Shamir (who is also the S in RSA). Let p be a large prime number. We identify the secret S with an element of the finite prime field $\mathbb{F}_p$. Similarly, we identify the n users of the scheme with n distinct nonzero elements $c_1, \ldots, c_n$ of $\mathbb{F}_p$. For this we must of course take $p > n$.

Algorithm 2.6.5 (Shamir Threshold Scheme) Let $n \geq 2$ be the number of users, let $p > n$ be a large prime number, let $c_1, \ldots, c_n \in \mathbb{F}_p^*$ be the user identifiers, let k be the threshold, and let $S \in \mathbb{F}_p$ be the secret. A trusted authority chooses random elements $a_1, \ldots, a_{k-1} \in \mathbb{F}_p$ and sets up the polynomial

$$f(x) = a_{k-1}x^{k-1} + \cdots + a_1 x + S \in \mathbb{F}_p[x]$$

of degree at most $k - 1$. The shares are obtained by $S_i = f(c_i)$ for $1 \leq i \leq n$ and then distributed to the users.

In order to prove that this is indeed a (k, n)-threshold scheme, we need to verify two properties: (i) any k function values of $f(x)$ determine $f(x)$ uniquely; (ii) $k-1$ function values of $f(x)$ do not determine $f(x)$. The second property is easy: if $f(b_1), \dots, f(b_{k-1})$ are given function values, then the polynomial

$$g(x) = f(x) + c(x - b_1) \cdots (x - b_{k-1}) \in \mathbb{F}_p[x]$$

with an arbitrary $c \in \mathbb{F}_p$ has degree at most $k-1$ and the same function values at $b_1, \dots, b_{k-1}$, that is, $g(b_j) = f(b_j)$ for $1 \le j \le k-1$. Note also that $S = f(0)$ and

$$g(0) = S + c(-b_1) \cdots (-b_{k-1}) = S + cd$$

for some nonzero $d \in \mathbb{F}_p$ if $b_1, \dots, b_{k-1}$ are nonzero, and so $g(0)$ contains no information about S since $S + cd$ runs through $\mathbb{F}_p$ if c runs through $\mathbb{F}_p$.

For the verification of the first property and the reconstruction of the secret, there are two methods that we can use. In the first method, we start from the given data $f(b_1), \dots, f(b_k)$ for distinct $b_1, \dots, b_k \in \mathbb{F}_p$, say $f(b_j) = f_j \in \mathbb{F}_p$ for $1 \le j \le k$. By writing down $f(x)$ in detail, we get

$$a_{k-1} b_j^{k-1} + \cdots + a_1 b_j + S = f_j \qquad \text{for } 1 \le j \le k.$$

This can be viewed as a system of k linear equations for the k unknowns $a_{k-1}, \dots, a_1, S$. The determinant D of the coefficient matrix is a Vandermonde determinant (we assume here that you are familiar with determinants), and by the well-known formula for Vandermonde determinants we obtain

$$D = \prod_{1 \le h < j \le k} (b_j - b_h).$$

Since $b_j - b_h \neq 0$ for $1 \le h < j \le k$, we have $D \neq 0$, and so the system of linear equations can be solved uniquely in $\mathbb{F}_p$.

In the second method, we are again given $f(b_j) = f_j \in \mathbb{F}_p$ for $1 \le j \le k$. Here $f(x)$ is explicitly computed by the Lagrange interpolation formula

$$f(x) = \sum_{j=1}^{k} f_j \prod_{\substack{h=1 \\ h \neq j}}^{k} (b_j - b_h)^{-1} (x - b_h).$$

The secret S is obtained from $S = f(0)$. It is easy to check that $f(b_j) = f_j$ for $1 \le j \le k$, since the above product over h is 1 at $x = b_j$ and 0 at $x = b_r$ for $r \neq j$. The uniqueness of $f(x)$ can be proved independently of the first method, for if $v(x) \in \mathbb{F}_p[x]$ is an arbitrary polynomial of degree $\le k-1$ with $v(b_j) = f_j$ for $1 \le j \le k$, then $(v - f)(b_j) = 0$ for $1 \le j \le k$. Thus, the polynomial $v(x) - f(x)$ of degree at most $k-1$ has k distinct roots $b_1, \dots, b_k$. But this is possible only if $v(x) - f(x)$ is the zero polynomial (see Theorem 1.4.27), and then $v(x) = f(x)$.

It is important to observe that the security of the Shamir threshold scheme does not rely on any unproved assumptions, unlike that of many other cryptographic schemes. In other words, the Shamir threshold scheme offers unconditional security.

The Shamir threshold scheme has several other nice properties. For instance, it is easy to add new users without changing the shares of the existing users. The trusted authority just chooses a nonzero identifier $c_{n+1} \in \mathbb{F}_p$ that has not been utilized before and assigns the share $S_{n+1} = f(c_{n+1})$ to the $(n+1)$st user. This does not affect the existing shares. Similarly, we can implement various levels of control. If the user Alice is higher up in the hierarchy, she can be provided with multiple shares corresponding to several different user identifiers. This gives more weight to Alice in coalitions of users.

On the other hand, the Shamir threshold scheme can be applied only once with a fixed set of shares. As soon as the members of a coalition of at least k users have disclosed their shares to recover the secret, these shares and the polynomial $f(x)$ are compromised. The trusted authority then has to choose a new random polynomial $f(x)$ and distributes new shares accordingly.

2.7 A glimpse of advanced topics

Although we presented several algorithms to solve the number-theoretic problems that form the basis of most public-key cryptosystems, namely the factorization problem for integers and the discrete logarithm problem, none of these algorithms is efficient enough to endanger, say, the RSA cryptosystem or the Diffie-Hellman key exchange if the parameters are carefully chosen. However, not much is known when we ask for the complexity of these problems in the rigorous sense of complexity theory in computer science. The book of Shparlinski [182] introduces new ways of using number theory in cryptography for the purpose of deriving lower bounds on the complexity of these number-theoretic problems. In particular, the book contains lower bounds on the degrees or orders of polynomials, algebraic functions, Boolean functions, and linear recurring sequences coinciding with the discrete logarithm for the finite prime field $\mathbb{F}_p$ at sufficiently many points. Just to whet the appetite, we state a sample result (see [182, Theorem 8.1]): let $f(x) \in \mathbb{F}_p[x]$, let g be a primitive element of $\mathbb{F}_p$, and let $S \subseteq \mathbb{F}_p^*$ be such that $\mathrm{ind}_g(a) = f(a)$ in $\mathbb{F}_p$ for all $a \in S$; then $\deg(f(x)) \geq 2|S| - p$, where we put $\deg(0) = 0$.

It is obvious that the definition of the discrete logarithm in Definition 2.4.1 makes sense in any finite cyclic group. The index-calculus algorithm (see Subsection 2.4.2) uses some special features of finite fields and is, with a proper choice of parameters, essentially faster than any so-called generic algorithm, that is, an algorithm such as the baby-step giant-step algorithm that can be easily extended from $\mathbb{F}_q^*$ to any finite cyclic group. This can be considered a disadvantage for cryptographic schemes based on the discrete logarithm problem for finite fields. Therefore cryptologists have looked around for finite cyclic groups, and more generally finite abelian groups, other than $\mathbb{F}_q^*$ that can be used as the basis for cryptographic schemes.

The points on an elliptic curve over a finite field form a finite abelian group for which the discrete logarithm problem is believed to be harder than the discrete logarithm problem for finite fields of a similar size because of the lack of an analog of the index-calculus algorithm for elliptic curves. Elliptic curves can be employed in versions of cryptographic schemes based on the discrete logarithm problem, as for example the Diffie-Hellman key exchange. We emphasize that elliptic curves are not ellipses. Elliptic curves received their name from integrals in calculus that arise in the computation of the arc length of ellipses.

An *elliptic curve* E over a finite field $\mathbb{F}_q$ of characteristic different from 2 and 3 is the set of solutions $(x, y) \in \mathbb{F}_q^2$ of a cubic polynomial equation of the form

$$y^2 = x^3 + ax + b, \quad a, b \in \mathbb{F}_q, \quad 4a^3 + 27b^2 \neq 0, \tag{2.1}$$

together with a further point O called the point at infinity. We turn E into a finite abelian group with the additive notation by stipulating first of all that O serves as the identity element of E. Next we describe how the inverse element $-P$ of a point P on E is defined. If $P = O$, then $-O = O$ by the rules for abelian groups. If $P = (x, y) \neq O$, then we put $-P = (x, -y)$. The axioms for abelian groups force us to define $P + O = O + P = P$ and $P + (-P) = (-P) + P = O$ for all points P on E. It remains to define the sum $P + Q$ for two points P and Q on E that are different from O and satisfy $Q \neq -P$. We first express the definition geometrically. We consider the line in $\mathbb{F}_q^2$ through P and Q if $P \neq Q$, and the tangent line to E at P if $P = Q$. This line intersects E in a unique third (respectively second) point R, and then by definition $P + Q = -R$. The arithmetic definition says that if $P \neq O$, $Q \neq O$, and $Q \neq -P$, then the sum of $P = (x_1, y_1)$ and $Q = (x_2, y_2)$ is given by $P + Q = (x_3, y_3)$ with

$$x_3 = c^2 - x_1 - x_2, \quad y_3 = c(x_1 - x_3) - y_1,$$

where

$$c = \begin{cases} (y_2 - y_1)(x_2 - x_1)^{-1} & \text{if } P \neq Q, \\ (3x_1^2 + a)(2y_1)^{-1} & \text{if } P = Q. \end{cases}$$

It requires some nontrivial computations to verify that E is indeed a finite abelian group under this addition (see the books on elliptic-curve cryptography cited below). The order $N(E)$ of E satisfies

$$|N(E) - q - 1| \leq 2q^{1/2}$$

according to the celebrated Hasse-Weil bound (see [47, Theorem 3.61] and [199, Theorem 4.2]).

Now we can describe the analog of the Diffie-Hellman key exchange (see Algorithm 2.4.3) for elliptic curves over a finite field. In a nutshell, we replace the large cyclic group $\mathbb{F}_q^*$ by a large cyclic subgroup of the finite abelian group E. We use the abbreviation $nP = \underbrace{P + P + \cdots + P}_{n \text{ summands}}$ for all positive integers n and all points P on E. All participants of a communication system share now a point P of large order t on an elliptic curve E over a finite field. In the first step, Alice chooses a random integer h with $2 \leq h \leq t-1$ and Bob chooses a random integer k with $2 \leq k \leq t-1$. In the second step, Alice sends hP to Bob over the channel, while Bob sends kP to Alice over the channel. The common key is $(hk)P$, which Alice computes as $h(kP)$ and Bob computes as $k(hP)$.

It should be obvious how to design, for example, an analog of the ElGamal cryptosystem (see Algorithm 2.4.4) in the framework of elliptic curves over finite fields. For more details on elliptic-curve cryptography, we refer first and foremost to the *Handbook of Elliptic and Hyperelliptic Curve Cryptography* [27]. There are also quite a number of monographs on this subject; we mention Blake, Seroussi, and Smart [12], Enge [47], Menezes [115], Washington [199], and two books by Koblitz [82], [83]. It is remarkable that elliptic curves can be used also for factoring integers and for primality tests (see [12, Chapter IX] and [82, Chapter VI]).

Groups derived from more complicated curves than elliptic curves have also been studied and may be attractive alternatives. For more information about this fascinating area, we refer again to the handbook [27] as well as to the books [12], [83], [147], and [148].

Stream ciphers represent an approach to symmetric encryption different from block ciphers. In a *stream cipher* the message is represented as a (usually finite) sequence $m_1, m_2, \ldots$ of bits and the message is encrypted by combining it with another (usually finite) sequence $k_1, k_2, \ldots$ of bits, the *keystream*. One possibility to encrypt the message is to view bits as elements of the finite field $\mathbb{F}_2$ and to add the message bits and the keystream bits term by term in $\mathbb{F}_2$. Thus, the ciphertext is the sequence $c_1, c_2, \ldots$ given by $c_i = m_i + k_i \in \mathbb{F}_2$ for all $i \geq 1$. The receiver can recover the message by adding the keystream to the ciphertext term by term, that is, by computing $m_i = c_i + k_i \in \mathbb{F}_2$ for all $i \geq 1$. This is extremely fast and has the nice feature that the same device can be used for encryption and decryption. In theory we could play the same game over any finite field, where in general decryption means subtracting the keystream from the ciphertext, but as so often in cryptology the Yogi Berra principle in Remark 2.4.5 applies again and motivates us to stick to $\mathbb{F}_2$.

If the keystream is a truly random sequence of bits, then we get the *one-time pad* or *Vernam cipher*, the latter named after the engineer Gilbert Vernam who got a U.S. patent for this cipher in 1919. The one-time pad is theoretically supported by the Shannon theorem (see [193, Chapter 2]) which says that the ciphertext in a one-time pad does not leak any information and so cannot be decrypted by an adversary. Thus, the one-time pad would be the Holy Grail of cryptography, but the trouble is that nobody knows how to generate a truly random sequence of bits in practice (see Sections 5.1 and 5.4 for discussions of this issue). Therefore the one-time pad is an idealization and stream ciphers attempt to imitate this ideal as best as they can. By the way, the name "one-time pad" stems from the requirement that the same keystream should not be reused to encrypt a second message. Indeed, if k_i is used to encrypt m_i and m_i', then by adding the identities $c_i = m_i + k_i$ and $c_i' = m_i' + k_i$ we obtain $c_i + c_i' = m_i + m_i'$. The adversary can compute $c_i + c_i'$ for as many values of i as desired, and this leaks information about m_i and m_i' for that many values of i.

Stream ciphers can be considered the practical versions of the one-time pad where the keystream is now a deterministically generated and often periodic sequence of bits with certain desirable features of randomness. Such a sequence is called a sequence of *pseudorandom bits*. We will say more about such sequences in Section 5.4. Keystreams generated by number-theoretic methods are discussed at length in the monograph [32]. Because of the difficulty of generating good keystreams and the practical problem of how to get the keystream from the sender to the receiver, stream ciphers are nowadays used only in contexts where a very high level of security is demanded and where a hierarchical organizational structure exists, for instance in military and diplomatic communications. The practicality of stream ciphers is also hampered by the fact that the keystream needs to be as long as the plaintext message, which causes problems when encrypting big data sets.

We mention a particularly elegant number-theoretic sequence of pseudorandom bits, where we again view bits as elements of $\mathbb{F}_2$. Let p be a large (and therefore odd) prime number and define $k_1, k_2, \ldots$ by $k_i = 1$ if i is a quadratic nonresidue modulo p and $k_i = 0$ otherwise. This sequence is periodic with least period length p. In a full period, there are $(p-1)/2$ terms equal to 1 and $(p+1)/2$ terms equal to 0 (compare with Remark 1.2.26). The sequence satisfies the obvious linear recurrence relation $k_{i+p} = k_i$ for all $i \geq 1$. This raises the interesting question of the least order of a linear recurrence relation that generates the sequence. This least order is called the *linear complexity* of a periodic sequence (see also Section 5.4). If L is the linear complexity of the given sequence $k_1, k_2, \ldots$, then there exist $c_0, c_1, \ldots, c_L \in \mathbb{F}_2$ with $c_L = 1$ such that

$$\sum_{l=0}^{L} c_l k_{i+l} = 0 \qquad \text{for all } i \geq 1.$$

If i is not divisible by p, then $(-1)^{k_i}$ is the Legendre symbol $\left(\frac{i}{p}\right)$, and so we obtain

$$\prod_{l=0}^{L} \left(\frac{i+l}{p}\right)^{c_l} = 1 \qquad \text{for } 1 \leq i \leq p - L - 1.$$

If we use the quadratic character η of $\mathbb{F}_p$ in Remark 1.4.53, then this can be written as

$$\eta\Big(\prod_{l=0}^{L}(i+l)^{c_l}\Big) = 1 \qquad \text{for } 1 \le i \le p-L-1.$$

Therefore we get

$$\begin{aligned} p-L-1 &= \sum_{i=1}^{p-L-1} \eta\Big(\prod_{l=0}^{L}(i+l)^{c_l}\Big) \\ &\le \Big|\sum_{i=0}^{p-1} \eta\Big(\prod_{l=0}^{L}(i+l)^{c_l}\Big)\Big| + L + 1 \le Lp^{1/2} + L + 1. \end{aligned}$$

In the last step we applied the Weil bound stated later in (6.8) in Subsection 6.1.3. We conclude that

$$Lp^{1/2} + 2(L+1) \ge p,$$

and so $L \ge \frac{1}{2}p^{1/2}$ for $p \ge 11$. Using a different method (see [32, Theorem 9.3.2]), one can determine the exact value of L, and in particular one obtains the lower bound $L \ge (p-1)/2$ for all odd prime numbers p. However, the method we have employed can be extended to produce lower bounds on the linear complexity for parts of the period (see [182, Theorem 9.2]), with an appropriate definition for the linear complexity of a finite sequence.

It is noteworthy that cryptosystems can also be utilized for the generation of keystreams. Let us start with DES. Take an arbitrary block M_1 of 64 bits. In the standard application of DES as a block cipher, M_1 would be a plaintext, but now M_1 is viewed as the initial value of a recursion. In detail, we recursively generate a sequence $M_1, M_2, \dots$ of 64-bit blocks by

$$M_{i+1} = \mathrm{DES}_K(M_i) \qquad \text{for } i = 1, 2, \dots,$$

where K is a fixed DES key. The sequence $M_1, M_2, \dots$ of 64-bit blocks is then regarded in an obvious manner as a sequence of bits and so as a keystream. We can play the same game with AES, but of course the blocks M_i consist then of 128 bits. It is difficult to carry out a theoretical analysis of the keystreams generated by DES and AES, but particularly for AES the keystreams perform satisfactorily under statistical tests for randomness. For the RSA cryptosystem, we again use the idea of repeated encryption. With the notation in Algorithm 2.3.5, we start from an initial value $m_1 \in Z_n$ and generate a sequence $m_1, m_2, \dots$ of elements of Z_n by the recursion

$$m_{i+1} \equiv m_i^e \pmod{n} \qquad \text{for } i = 1, 2, \dots.$$

Trivial initial values such as $m_1 = 0, 1, n-1$ have to be excluded. A sequence $k_1, k_2, \dots$ of bits is obtained by the formula

$$k_i \equiv m_i \pmod{2} \qquad \text{for } i = 1, 2, \dots,$$

and this is a keystream produced by the RSA cryptosystem.

Cryptography is such a wide area that a lot more can be said about it, but we have to respect certain limits. The books cited in Subsection 2.1.1 will certainly satisfy your curiosity. A few more topics related to cryptography will be discussed in the following chapters, including code-based cryptosystems (see Section 3.6) and the possible impact of quantum computers on cryptography (see Subsection 6.5.1).

Exercises

2.1 Use the decryption function D given in Example 2.1.2 to decrypt YCDLMF ZV PUS.

2.2 Consider the affine cipher determined by $e(m) \equiv 7m + 12 \pmod{26}$. We identify the English alphabet with Z_{26} by $A \leftrightarrow 0,\ B \leftrightarrow 1, \ldots, Z \leftrightarrow 25$.

(a) Encrypt the word BECKENBAUER using this substitution.

(b) Which word was encrypted to 4 1 12 25 4 11? (Hint: the solution is the name of an Austrian football player who would have become less famous if Franz Beckenbauer had attended the World Cup 1978.)

2.3 For a linear substitution $e(m) \equiv am + b \pmod{31}$ with $a, b \in Z_{31}$, we know $e(2) = 5$ and $e(3) = 10$.

(a) Determine a and b.

(b) Determine the inverse map e^{-1}.

2.4 Verify that the permutations π and π^{-1} used in DES satisfy $\pi^{-1}(\pi(i)) = i$ for all $i = 1, \ldots, 64$.

2.5 Verify that all steps in AES are invertible.

2.6 Prove the assertions in Remark 2.2.3 in detail.

2.7 Show that $f(x) = x^3$ is an APN function over every finite field $\mathbb{F}_{2^r}$ with $r \in \mathbb{N}$.

2.8 (a) Encrypt the messages $m = 5$ and $m = 7$ with the RSA cryptosystem and the public key $(n, e) = (35, 5)$.

(b) Calculate the private key d.

2.9 Determine all possible encryption exponents $e \leq 60$ for the RSA modulus $n = 77$.

2.10 Determine the number of multiplications modulo n that are required for an RSA encryption with modulus n and encryption exponent $e = 2^{18} + 2^8 + 1$ if an efficient algorithm is used for this purpose.

2.11 A plaintext $m \in Z_n$ in the RSA cryptosystem with public key (n, e) is said to be fixed if $m^e \equiv m \pmod{n}$. Prove that the number of fixed plaintexts is given by

$$(\gcd(p-1, e-1)+1)(\gcd(q-1, e-1)+1).$$

2.12 Show that the encryption exponent $e = \phi(n)/2 + 1$ is unsuitable in the RSA cryptosystem since then $m^e \equiv m \pmod{n}$ for all $m \in \mathbb{Z}$.

2.13 (a) Suppose that the same plaintext $m \in Z_n$ is encrypted twice with the RSA cryptosystem using two public keys (n, e) and (n, f) with $\gcd(e, f) = 1$. Show that m can be recovered from the two ciphertexts $c_e \equiv m^e \pmod{n}$ and $c_f \equiv m^f \pmod{n}$.

(b) Consider the special case $n = 77$, $e = 13$, and $f = 17$. The two ciphertexts are $c_e = 3$ and $c_f = 5$. Find the plaintext m.

2.14 Consider the RSA cryptosystem with public key (n, e).

(a) Prove that there exists a positive integer k such that $m^{e^k} \equiv m \pmod{n}$ for all $m \in \mathbb{Z}$.

(b) For an integer k in part (a), prove that $c^{e^{k-1}} \equiv m \pmod{n}$ for the ciphertext c corresponding to the plaintext m.

(c) Suppose that a small integer k with the property in part (a) can be found. Argue that this endangers the security of this RSA cryptosystem.

2.15 Does an analog of the RSA cryptosystem work if n is the product of more than two distinct prime numbers? What is the disadvantage if we take more than two prime factors? Why is $n = p^2$ a bad choice?

2.16 Suppose that $m \in \mathbb{N}$ is divisible by the square of a prime number. Prove that there exist integers a_1 and a_2 such that $a_1 \not\equiv a_2 \pmod{m}$, but $a_1^k \equiv a_2^k \pmod{m}$ for all integers $k \geq 2$.

2.17 The *Rabin cryptosystem* works with a modulus $n = pq$, where p and q are distinct prime numbers with $p \equiv 3 \pmod{4}$ and $q \equiv 3 \pmod{4}$. Furthermore, an integer b with $0 \leq b \leq n-1$ is chosen. The public key is the ordered pair (n, b), while p and q form the private key. A plaintext $m \in Z_n$ is encrypted by computing $c \in Z_n$ with

$$c \equiv m(m+2b) \pmod{n}.$$

(a) Show that the least residues of $w(m+b) - b$ modulo n encrypt to the same ciphertext, where w is any of the four solutions of $x^2 \equiv 1 \pmod{n}$ provided by Lemma 2.3.15. This is a rare example of an encryption function which is not injective.

(b) For a prime number $p \equiv 3 \pmod 4$ and a quadratic residue a modulo p, show that $x \equiv a^{(p+1)/4} \pmod p$ is a solution of $x^2 \equiv a \pmod p$.

(c) In order to decrypt a Rabin ciphertext c, the quadratic congruence $x^2 + 2bx \equiv c \pmod n$ has to be solved. By standard substitutions from the theory of quadratic equations, this is equivalent to solving $x^2 \equiv a \pmod n$ for some $a \in \mathbb{Z}$. Show that the latter congruence can be solved in a straightforward manner by the legitimate receiver if the encryption was performed correctly. (Note: Among the up to four possible plaintexts, the one which is most plausible is taken as the correct one. The sender may also deliberately create redundancy in the plaintext to facilitate decryption.)

2.18 Try to factor 1927, 7721, 11413, 17111, and 200819 using:

(a) trial division, that is, checking whether one of the first prime numbers $2, 3, 5, 7, 11, \ldots$ is a divisor;

(b) Fermat factorization;

(c) the Pollard rho algorithm;

(d) square-root factoring.

2.19 Find all solutions $x \in Z_{105}$ of the congruence $x^2 \equiv 16 \pmod{105}$. (Note that Lemma 2.3.15 does not apply since 105 has three distinct prime factors.)

2.20 Let $f : S \to S$ be a self-map of a set S and let $s_0, s_1, \ldots$ be a sequence of elements of S generated by $s_i = f(s_{i-1})$ for $i = 1, 2, \ldots$ with an arbitrary initial value s_0. Suppose that $s_0, s_1, \ldots, s_{15}$ are distinct, but that $s_{16} = s_9$. Find the least integer $\ell \geq 1$ with $s_\ell = s_{2\ell}$.

2.21 For $q = 53$, $g = 2$, $h = 29$, and $k = 19$, describe the Diffie-Hellman key exchange. Work out the common key of Alice and Bob.

2.22 Let $\mathbb{F}_q$ be a finite field and let g be a primitive element of $\mathbb{F}_q$. Prove that

$$\mathrm{ind}_g(-a) \equiv \mathrm{ind}_g(a) + \frac{q-1}{2} \pmod{q-1} \qquad \text{for all } a \in \mathbb{F}_q^*.$$

2.23 Let $\mathbb{F}_q$ be a finite field and let g and h be primitive elements of $\mathbb{F}_q$. Prove that

$$\mathrm{ind}_h(a) \equiv \mathrm{ind}_g(a) \cdot \mathrm{ind}_h(g) \pmod{q-1} \qquad \text{for all } a \in \mathbb{F}_q^*.$$

2.24 Compute the discrete logarithm modulo 113 of $a = 57$ to the base 3 using the baby-step giant-step algorithm.

2.25 Compute the discrete logarithm modulo 29 of $a = 18$ to the base 2 using the Silver-Pohlig-Hellman algorithm.

2.26 Compute the discrete logarithm modulo 229 of $a = 13$ to the base 6 using the index-calculus algorithm. (Hint: choose $B = 11$.)

2.27 Use the data in Example 2.4.8 to solve the power congruence $3^k \equiv 6 \pmod{19}$ for the integer k.

2.28 The public key of Bob in an ElGamal signature scheme is $(p, g, a) = (107, 2, 80)$. He signs his message with $(b, c) = (9, 93)$. Show that the message $m = 17$ can be sent by him with this signature being accepted as valid, but that $m = 10$ and $m = 83$ are forged.

2.29 Suppose that Bob is using the ElGamal signature scheme and that he signs two plaintexts m_1 and m_2 with signatures (b, c_1) and (b, c_2), respectively, so that the same value of b occurs in the first entry of both signatures. Suppose also that $\gcd(c_1 - c_2, p - 1) = 1$.

(a) Describe how r can be computed efficiently given this information.

(b) Show that the signature scheme can then be broken.

2.30 Suppose that in a Shamir threshold scheme the parameters are $p = 19$, $k = 3$, and $n = 6$, and shares are given by $f(2) = 8$, $f(3) = 18$, and $f(6) = 11$. Calculate the secret $S = f(0)$.

2.31 (a) Verify that $c^{2ab} = c^{(a+b)^2} c^{-a^2} c^{-b^2}$ for all $c \in \mathbb{F}_q^*$ and $a, b \in \mathbb{N}$.

(b) Note that 2 is a quadratic residue modulo p if and only if $p \equiv \pm 1 \pmod 8$ (see [152, Theorem 3.3]). For a prime number $p \equiv 5 \pmod 8$ and a quadratic residue a modulo p, show that

$$x \equiv \begin{cases} a^{(p+3)/8} \pmod p & \text{if } a^{(p-1)/4} \equiv 1 \pmod p, \\ \frac{p+1}{2}(4a)^{(p+3)/8} \pmod p & \text{if } a^{(p-1)/4} \equiv -1 \pmod p, \end{cases}$$

is a solution of $x^2 \equiv a \pmod p$. (Note: since square-root finding in finite fields is an easy task, the Diffie-Hellman map $D(g^a, g^b) \equiv g^{ab} \pmod p$ could be efficiently evaluated if the univariate map $d(g^a) \equiv g^{a^2} \pmod p$ can be represented by a low-degree polynomial.)

(c) Let p be a prime number and let $f(x) \in \mathbb{F}_p[x]$ with $f(g^a) = g^{a^2}$ in $\mathbb{F}_p$ for all $a \in S \subseteq \{0, 1, \ldots, p-2\}$, where g is a primitive element of $\mathbb{F}_p$. Prove a lower bound on the degree of $f(x)$ in terms of p and $|S|$. (Hint: estimate the number of roots of the polynomial $m(x) := f(gx) - gx^2 f(x)$ with $\deg(m(x)) = \deg(f(x)) + 2$.)

2.32 Let F be a field. Show that the three roots of $x^3 + ax + b \in F[x]$ are distinct if and only if $4a^3 + 27b^2 \neq 0$.

2.33 Prove that the number of ordered pairs $(a, b) \in \mathbb{F}_q^2$ with $4a^3 + 27b^2 \neq 0$ is equal to $q^2 - q$.

2.34 Consider the elliptic curve E over $\mathbb{F}_{11}$ defined by $y^2 = x^3 + 8x$. Show that $P = (8, 9)$ and $Q = (9, 3)$ are points on E and compute $P + Q$ and $2P$.

2.35 Consider the elliptic curve E over $\mathbb{F}_7$ defined by $y^2 = x^3 + 5x + 4$. Show that E is a cyclic group. (Hint: determine the 10 points on the curve and show that $P = (3, 2)$ is a point of order 10.)

2.36 Let E be the elliptic curve over $\mathbb{F}_q$ given by (2.1) and let η be the quadratic character of $\mathbb{F}_q$ with $\eta(d) = 0$ for $d = 0 \in \mathbb{F}_q$. Prove that the order $N(E)$ of E is given by

$$N(E) = q + 1 + \sum_{c \in \mathbb{F}_q} \eta(c^3 + ac + b).$$

2.37 Describe the analog of the ElGamal cryptosystem for elliptic curves.

2.38 Let g be a primitive root modulo a prime number $p > 2$ and consider the periodic sequence $k_1, k_2, \ldots$ of elements of $\mathbb{F}_2$ with period length $p - 1$ that is defined by $k_i = 1$ if and only if $g^i + 1$ is a quadratic residue modulo p. Prove a lower bound on the linear complexity of this sequence.

Chapter 3

Coding Theory

If you select the codes of Reed
and Solomon or kindred breed
with shrewdness and not badly,
you will be coding gladly
as they meet every need.

3.1 Introduction to error-correcting codes

3.1.1 Basic definitions

Life is a comedy of errors, at least in the opinion of William Shakespeare, but you can make a concentrated effort to reduce the number of errors that you commit and thus increase the quality of your life. There is probably no panacea for all human errors and mishaps, but in the setting of communication technology, number theory and finite fields can help to prevent errors and ensure the quality of communication. The aim of this chapter is to explain in sufficient detail how this is achieved.

We consider the transmission of information through a communication medium, and as in Chapter 2 we use the convenient term *channel* for a communication medium. A channel can, for instance, be a computer network, a satellite link, the Internet, or even the interface between a storage medium (like a compact disk) and its reading device. In practice, channels are subject to various types of disturbance, distortion, and interference. This may cause transmission errors, and so the information that is received may not coincide with the information that was sent. Engineers speak of a *noisy channel* to designate a channel that may produce transmission errors. The frequency of transmission errors depends on the physical nature of the channel. For instance, one expects that in communications over very long distances, such as they occur in space missions, the error probability will be rather high.

A fundamental requirement in modern communication systems is reliability, meaning that information is received as sent. Reliability does not come for free. Indeed, special features and algorithms have to be built into a communication system to guarantee that transmission errors are eliminated. This is exactly where coding schemes and coding theory enter the scene. In simple terms, a *coding scheme* is an algorithm and/or a device for detecting and correcting transmission errors that occur in noisy channels. At the core of a coding scheme is the mathematical concept of an *error-correcting code*, or simply a *code*. Coding schemes are nowadays omnipresent in communication systems and also in storage systems. They are normally fully integrated into these systems, and then the user is actually not aware that error control is taking place.

In practice, channels are both noisy and insecure. But the protection against noisy channels and the protection against insecure channels are different as mathematical problems, and so it is customary to treat these issues separately. We also follow this tradition, and therefore we discussed the protection against insecure channels, that is cryptography, in Chapter 2 and we deal with the protection against noisy channels, that is error-correcting codes, in the present chapter. In the real world, security and reliability have to be provided concurrently in our communication.

Example 3.1.1 The message DAD SEND MONEY in Example 2.1.2 is now sent over a noisy channel. If the noisy channel is in addition a malicious channel, then it may deliver DAD SEND HONEY to the recipient, thus causing considerable confusion. The lesson is again that important messages should not only be encrypted so as to frustrate eavesdroppers, but should also be protected against transmission errors by using an error-correcting code.

Coding theory (that is, the theory of coding schemes) is a wide subject at the interface of discrete mathematics and information theory. One may therefore distinguish between the part of coding theory oriented more towards discrete and structural mathematics (this part is often called algebraic coding theory) and the information-theoretic part which studies channels from a probabilistic viewpoint. Both parts are covered very well in the book of McEliece [113].

The history of coding theory has a well-marked beginning with the seminal paper of Shannon [179] from 1948 which introduced the basic information-theoretic model for coding theory and established fundamental existence results. Claude Shannon (1916–2001) was a brilliant mathematician and also a quirky character who liked to ride a unicycle in the halls of the AT&T Bell Laboratories at night. He built not only coding theory, but also juggling machines and one of the first chess computers.

Shortly after the publication of Shannon's paper [179], various explicit error-correcting codes were constructed, some of which still belong to the standard repertoire of coding theory. We will meet these classical codes, such as the Hamming codes and the Golay codes, later in this chapter. The 1950s and 1960s saw dramatic progress in coding theory, so that by the end of the latter decade coding theory was already a rich and well-founded subject. Significantly, the very influential monograph of Berlekamp [10] on algebraic coding theory appeared near the end of the 1960s. Other milestones in the expository literature on coding theory are the book of MacWilliams and Sloane [108] and the *Handbook of Coding Theory* edited by Pless and Huffman [162].

In line with the general perspective of this book, we focus on the number-theoretic aspects of coding theory. Number theory plays indeed a major role in the construction of efficient error-correcting codes. The basic structure for this purpose is that of a finite field (see Section 1.4). Besides finite fields, we use also elementary linear algebra and simple facts about rings and ideals. With these tools, the coverage of algebraic coding theory can be pushed quite far, and so we will be able to treat the fundamentals of algebraic coding theory and the most important specific codes in this chapter.

In order to formalize coding schemes, we start by considering the data to be transmitted. We assume that these data are formatted as a string of symbols from a chosen alphabet. Since modern communication is digital, it is reasonable to select a finite alphabet. Frequently, the alphabet consists of the bits 0 and 1, but sometimes it is more efficient to use an alphabet of larger size. From the theoretical point of view, it is preferable to choose an alphabet with mathematical structure. In fact, we assume that the alphabet is a finite field $\mathbb{F}_q$ of order q for some prime power q. Thus, the data are formatted as a string of elements of $\mathbb{F}_q$. The next step in the preparation for coding is to split up this string into blocks of fixed length, let us say of length $k \geq 1$. Some padding (say by zero elements) may be needed at the end to arrive at a partition into complete blocks of length k. The coding scheme now processes the data block by block.

From now on, we will thus assume that the input of the coding scheme is a block of length k of elements of $\mathbb{F}_q$, or in other words a k-tuple $(a_1, \ldots, a_k)$ with $a_i \in \mathbb{F}_q$ for $i = 1, \ldots, k$. We use the standard notation $\mathbb{F}_q^k$ for the set of all these k-tuples. An element of $\mathbb{F}_q^k$ is also called a *word* (over $\mathbb{F}_q$) of length k. The essential idea of a coding scheme is to take an input $\mathbf{a} = (a_1, \ldots, a_k) \in \mathbb{F}_q^k$ and add redundant information to allow for error correction. We assume that this transforms $\mathbf{a}$ into an n-tuple $\mathbf{c} \in \mathbb{F}_q^n$ with $n \geq k$. In fact, in nontrivial situations we suppose that $n > k$.

Example 3.1.2 Let $k = 1$ and let n be of the form $n = 2r + 1$ for some integer $r \geq 1$. An input block of the coding scheme consists thus of a single element $a \in \mathbb{F}_q$. We create redundancy by repeating this element n times. In other words, we set up the map

$$\psi : a \in \mathbb{F}_q \mapsto (a, \ldots, a) \in \mathbb{F}_q^n.$$

We send $\mathbf{c} = \psi(a)$ over the noisy channel. Assume that at most r errors can occur in this transmission. The receiver will then get an n-tuple $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{F}_q^n$ where at least $n - r = r + 1$ coordinates v_j are equal to a. Hence by looking at the coordinates of $\mathbf{v}$ and observing that a is the most frequent one, the receiver can recover the correct element $a \in \mathbb{F}_q$ although up to r errors may have occurred in the transmission. For instance, if $r = 2$, $n = 5$, and the quintuple $\mathbf{v} = (1, 0, 1, 1, 0) \in \mathbb{F}_q^5$ is received, then $a = 1$ since 1 is the most frequent coordinate of $\mathbf{v}$. This is a very simple scheme, and its shortcoming is that it incurs data expansion by a factor $n \geq 3$ and thus a corresponding loss of speed in the data transmission. The task of coding theory is to design more efficient schemes.

In general, the passage from the input $\mathbf{a} \in \mathbb{F}_q^k$ (also called the *message*) to $\mathbf{c} \in \mathbb{F}_q^n$ is described by an injective map $\psi : S \to \mathbb{F}_q^n$ from some nonempty subset S of $\mathbb{F}_q^k$ to $\mathbb{F}_q^n$. The image of ψ is called a code, and this leads to the following simple formal definition.

Definition 3.1.3 A *code* (over $\mathbb{F}_q$) is a nonempty subset C of $\mathbb{F}_q^n$. The integer $n \geq 1$ is called the *length* of the code C. An element of C is called a *codeword* in C.

With this terminology, we thus have an injective map ψ which takes a message $\mathbf{a} \in S \subseteq \mathbb{F}_q^k$ to a codeword $\mathbf{c} = \psi(\mathbf{a}) \in \mathbb{F}_q^n$, where normally $n > k$. The map ψ is called an *encoder*.

Example 3.1.4 Consider the encoder ψ in Example 3.1.2. The corresponding code is

$$C = \{(a, \ldots, a) \in \mathbb{F}_q^n : a \in \mathbb{F}_q\}$$

and its length is $n = 2r + 1$. As we have seen in Example 3.1.2, this code can correct up to $r = (n - 1)/2$ errors in a word of length n. For obvious reasons, C is called a *repetition code*.

A code over $\mathbb{F}_2$ is also called a *binary code* and a code over $\mathbb{F}_3$ is also called a *ternary code*. Similarly, one may speak of a *quaternary code* for a code over $\mathbb{F}_4$, and so on, and in general a code over $\mathbb{F}_q$ is sometimes referred to as a *q-ary code*.

3.1.2 Error correction

A primary characteristic of a code is its error-correction capability, that is, the number of errors that it can correct in a word of length n, where n is the length of the code. If we think of $\mathbf{c} \in \mathbb{F}_q^n$ as a sent word and $\mathbf{v} \in \mathbb{F}_q^n$ as a received word, then the number of errors is equal to the number of coordinates in which $\mathbf{c}$ and $\mathbf{v}$ differ. Thus, the following notion is highly relevant in this context.

Definition 3.1.5 For $\mathbf{c} = (c_1, \ldots, c_n) \in \mathbb{F}_q^n$ and $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{F}_q^n$, the *Hamming distance* $d(\mathbf{c}, \mathbf{v})$ is defined to be the number of coordinates in which $\mathbf{c}$ and $\mathbf{v}$ differ, that is,

$$d(\mathbf{c}, \mathbf{v}) = \# \{1 \leq j \leq n : c_j \neq v_j\}.$$

The Hamming distance is defined for pairs of words over $\mathbb{F}_q$ of any (but equal) length, and so in particular for pairs of words of length 1. If, for the moment, we let d_1 denote the Hamming distance for pairs of words of length 1, then with the notation in Definition 3.1.5 we can write

$$d(\mathbf{c}, \mathbf{v}) = d_1(c_1, v_1) + \cdots + d_1(c_n, v_n). \tag{3.1}$$

Note that for $c, v \in \mathbb{F}_q$ we have $d_1(c, v) = 0$ if $c = v$ and $d_1(c, v) = 1$ if $c \neq v$.

Proposition 3.1.6 *The Hamming distance d has the following properties for all $\mathbf{c}, \mathbf{u}, \mathbf{v} \in \mathbb{F}_q^n$:*
(i) $0 \leq d(\mathbf{c}, \mathbf{u}) \leq n$ (nonnegativity and upper bound)*;*
(ii) $d(\mathbf{c}, \mathbf{u}) = 0$ *if and only if* $\mathbf{c} = \mathbf{u}$ (identity of indiscernibles)*;*
(iii) $d(\mathbf{c}, \mathbf{u}) = d(\mathbf{u}, \mathbf{c})$ (symmetry)*;*
(iv) $d(\mathbf{c}, \mathbf{v}) \leq d(\mathbf{c}, \mathbf{u}) + d(\mathbf{u}, \mathbf{v})$ (triangle inequality).

Proof. The properties (i), (ii), and (iii) are trivial. In view of (3.1), it suffices to prove (iv) for $n = 1$, that is, for the Hamming distance d_1. We take $c, u, v \in \mathbb{F}_q$ and distinguish two cases. If $c = v$, then (iv) is obvious since $d_1(c, v) = 0$. If $c \neq v$, then either $c \neq u$ or $u \neq v$, and (iv) is again true for d_1. □

Example 3.1.7 Let $q = 2$ and let

$$\mathbf{c} = (1, 1, 1, 0, 1) \in \mathbb{F}_2^5, \ \mathbf{u} = (0, 1, 0, 0, 1) \in \mathbb{F}_2^5, \ \mathbf{v} = (0, 1, 0, 1, 0) \in \mathbb{F}_2^5.$$

Then $d(\mathbf{c}, \mathbf{u}) = 2$, $d(\mathbf{c}, \mathbf{v}) = 4$, $d(\mathbf{u}, \mathbf{v}) = 2$, and so (iv) in Proposition 3.1.6 is verified immediately. This example demonstrates that equality in (iv) can occur in a nontrivial situation.

Remark 3.1.8 If you had a course on analysis or topology, then you will recognize (ii), (iii), and (iv) in Proposition 3.1.6, together with the first inequality in (i), as the axioms for a distance function in a metric space. Thus, Proposition 3.1.6 shows that the pair $(\mathbb{F}_q^n, d)$ forms a metric space. This metric space is called a *Hamming space*.

Definition 3.1.9 For a code C containing at least two codewords, the *minimum distance* $d(C)$ of C is defined by

$$d(C) = \min \{d(\mathbf{c}_1, \mathbf{c}_2) : \mathbf{c}_1, \, \mathbf{c}_2 \in C, \ \mathbf{c}_1 \neq \mathbf{c}_2\}.$$

In words, $d(C)$ is the closest that two distinct codewords in C can come together in terms of the Hamming distance.

Example 3.1.10 Let $q = 2$ and let the binary code C of length 5 consist of the codewords

$$\mathbf{c}_1 = (0,0,0,0,0), \ \mathbf{c}_2 = (1,1,0,0,0), \ \mathbf{c}_3 = (1,1,1,1,1).$$

Then $d(\mathbf{c}_1, \mathbf{c}_2) = 2$, $d(\mathbf{c}_1, \mathbf{c}_3) = 5$, $d(\mathbf{c}_2, \mathbf{c}_3) = 3$, and so $d(C) = 2$.

Example 3.1.11 Consider the repetition code in Example 3.1.4. Then any two distinct codewords in C differ in all n coordinates, and so $d(C) = n$.

The minimum distance is a crucial parameter of a code since it governs the error-correction capability of a code, as we shall see in Theorem 3.1.14 below.

Let us now take a closer look at the issue of error correction. We recall that in our model for communication we start with a word $\mathbf{a}$ over $\mathbb{F}_q$ of length k (the message) and transform it by the encoder into a codeword $\mathbf{c}$ over $\mathbb{F}_q$ of length n, where typically $n > k$. The codeword $\mathbf{c}$ is sent over the noisy channel. On the other side of the channel we get a received word $\mathbf{v}$ over $\mathbb{F}_q$ of length n which may be different from $\mathbf{c}$. The problem of error correction is how to recover $\mathbf{c}$ from $\mathbf{v}$, if this is at all possible. The following diagram represents the general situation in our model and introduces two more devices, the error processor and the decoder.

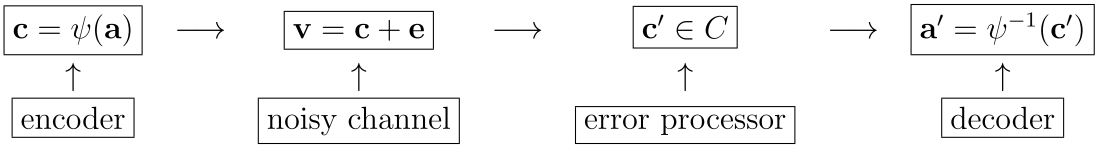
$\mathbf{c} = \psi(\mathbf{a})$
$\mathbf{v} = \mathbf{c} + \mathbf{e}$
$\mathbf{c}' \in C$
$\mathbf{a}' = \psi^{-1}(\mathbf{c}')$
encoder
noisy channel
error processor
decoder

Figure 3.1: The model for error correction

The *error processor* serves the purpose of error correction. It takes the input $\mathbf{v}$ and attempts to find the most likely codeword $\mathbf{c}'$ corresponding to it, by applying what is called a *decoding algorithm*. Again, $\mathbf{c}'$ may be different from the original codeword $\mathbf{c}$. The *decoder* applies the inverse map ψ^{-1} of the encoder ψ to $\mathbf{c}'$ and produces the output $\mathbf{a}' = \psi^{-1}(\mathbf{c}')$. In the case of a successful communication, we should have $\mathbf{c}' = \mathbf{c}$ and $\mathbf{a}' = \mathbf{a}$. The decoder raises no fundamental issues as it simply requires the application of the inverse map of a given injective map. On the other hand, the design of efficient decoding algorithms is a nontrivial problem which has received a lot of attention in coding theory.

An important mathematical task for the model above is to find sufficient conditions for a successful communication. The following terminology is convenient in connection with this task.

Definition 3.1.12 For an integer $t \geq 0$, a code $C \subseteq \mathbb{F}_q^n$ is called *t-error-correcting* if for every $\mathbf{v} \in \mathbb{F}_q^n$ there is at most one $\mathbf{c} \in C$ such that $d(\mathbf{v}, \mathbf{c}) \leq t$.

Every code is trivially 0-error-correcting, and so the concept in Definition 3.1.12 is of practical interest only for $t \geq 1$.

The standard principle used by the error processor is *nearest neighbor decoding*. It is based on the admittedly optimistic assumption that few rather than many errors occur in the transmission over the noisy channel.

Algorithm 3.1.13 (Nearest Neighbor Decoding) Let C be a code over $\mathbb{F}_q$ of length n.
Input: a received word $\mathbf{v} \in \mathbb{F}_q^n$.
Output: a codeword $\mathbf{c}' \in C$ that is closest to $\mathbf{v}$ in terms of the Hamming distance, that is,

$$d(\mathbf{v}, \mathbf{c}') = \min_{\mathbf{c} \in C} d(\mathbf{v}, \mathbf{c}).$$

The actual procedure of passing from the input $\mathbf{v}$ to the output $\mathbf{c}'$ depends on the nature of the code C, and more will be said about this later in this chapter. This procedure is the core of decoding algorithms, and from a practical point of view it is an essential requirement that it be reasonably efficient.

If $\mathbf{c} \in C$ is sent over the noisy channel and at most t transmission errors occur in this word of length n, then $d(\mathbf{v}, \mathbf{c}) \leq t$ for the received word $\mathbf{v}$. If we know that C is t-error-correcting, then $d(\mathbf{v}, \mathbf{z}) > t$ for all other codewords $\mathbf{z} \neq \mathbf{c}$ in C, which means that $\mathbf{c}$ is closest to $\mathbf{v}$ (in terms of the Hamming distance) among all codewords in C and nearest neighbor decoding gives the correct result. The missing link is the following result.

Theorem 3.1.14 *If C is a code with at least two codewords and with minimum distance $d(C)$, then C is t-error-correcting with $t = \lfloor (d(C) - 1)/2 \rfloor$.*

Proof. We proceed by contradiction. Let $C \subseteq \mathbb{F}_q^n$ and suppose that, for some $\mathbf{v} \in \mathbb{F}_q^n$, there exist two codewords $\mathbf{c}_1, \mathbf{c}_2 \in C$ with $\mathbf{c}_1 \neq \mathbf{c}_2$ such that $d(\mathbf{v}, \mathbf{c}_1) \leq t$ and $d(\mathbf{v}, \mathbf{c}_2) \leq t$. Then the triangle inequality yields

$$d(\mathbf{c}_1, \mathbf{c}_2) \leq d(\mathbf{c}_1, \mathbf{v}) + d(\mathbf{v}, \mathbf{c}_2) \leq 2t \leq d(C) - 1,$$

a contradiction to the definition of $d(C)$. □

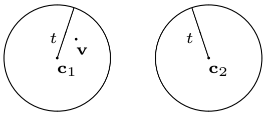

Figure 3.2: Unique error correction

Example 3.1.15 Consider again the repetition code C in Examples 3.1.2, 3.1.4, and 3.1.11. We noted in Example 3.1.11 that $d(C) = n = 2r + 1$, and so Theorem 3.1.14 implies that C is r-error-correcting. This agrees with the result of the simple analysis we carried out in Example 3.1.2.

Example 3.1.16 Let C be the binary code of length 5 consisting of the codewords

$$\mathbf{c}_1 = (0,0,0,0,0),\ \mathbf{c}_2 = (0,0,1,1,1),\ \mathbf{c}_3 = (1,1,0,1,1),\ \mathbf{c}_4 = (1,1,1,0,0).$$

It is straightforward to verify that $d(C) = 3$. Therefore C is 1-error-correcting by Theorem 3.1.14.

Remark 3.1.17 A simpler problem than error correction is error detection, where we want to recognize by looking at the received word $\mathbf{v}$ whether transmission errors have occurred or not. For an integer $u \geq 1$, a code $C \subseteq \mathbb{F}_q^n$ is called *u-error-detecting* if the property $1 \leq d(\mathbf{v}, \mathbf{c}) \leq u$ with $\mathbf{v} \in \mathbb{F}_q^n$ and $\mathbf{c} \in C$ always implies that $\mathbf{v} \notin C$. For such a code C, if $\mathbf{c} \in C$ is sent over the noisy channel and at most u transmission errors occur in this word of length n, then there are two possible cases for the received word $\mathbf{v}$: either (i) $\mathbf{v} \in C$, then $\mathbf{v} = \mathbf{c}$ and the transmission is error-free; or (ii) $\mathbf{v} \notin C$, then we have detected that transmission errors have happened. To decide whether we are in case (i) or in case (ii) is a simple matter of going through the list of codewords in C, and there may even be more efficient ways of deciding this in situations where C has a nice structural description. If C has at least two codewords and minimum distance $d(C) \geq 2$, then it is clear that C is u-error-detecting with $u = d(C) - 1$, for if $\mathbf{v} \in \mathbb{F}_q^n$, $\mathbf{c} \in C$, and $1 \leq d(\mathbf{v}, \mathbf{c}) \leq u = d(C) - 1$, then $\mathbf{v} \notin C$ by the definition of $d(C)$. In the remaining part of this chapter, we will interpret results about the minimum distance of a code in terms of the error-correction capability of the code, but the discussion above shows that an interpretation in terms of the error-detection capability of the code is possible as well.

We summarize the desirable properties of a good code and equivalently of a good coding scheme:

(i) large minimum distance of the code to guarantee a high error-correction capability (see Theorem 3.1.14);

(ii) not too much loss of speed in the data transmission caused by the code (a negative example is the repetition code in Example 3.1.2 with large n);

(iii) the computational procedures in the coding scheme (that is, the encoder, the decoding algorithm, and the decoder) are fast.

The goals (i) and (ii) are usually not compatible, as we will see in Subsection 3.4.2. Therefore, in general one has to settle for a trade-off between (i) and (ii).

3.2 Linear codes

3.2.1 Vector spaces over finite fields

The Hamming space $\mathbb{F}_q^n$ in Remark 3.1.8 can be endowed with additional structure, namely that of a vector space. You have learned about vector spaces in a course on linear algebra, but most likely you have seen only vector spaces over the real numbers and over the complex numbers in that course. In abstract linear algebra one can consider vector spaces over any field, and so in particular over a finite field, which is the relevant case for the theory of linear codes. The theory of vector spaces works basically in the same way for any field of scalars. For your benefit, we briefly review the fundamentals of vector spaces over finite fields.

A *vector space* (or *linear space*) V over $\mathbb{F}_q$ has two operations: addition of vectors from V and multiplication of a vector from V by a scalar from $\mathbb{F}_q$. These operations have to satisfy certain properties. To begin with, V is an abelian group with respect to vector addition. The identity element of this group is called the *zero vector* and denoted by $\mathbf{0}$. Multiplication by scalars is distributive with regard to both vector addition and scalar addition, that is, $c(\mathbf{a}+\mathbf{b}) = c\mathbf{a}+c\mathbf{b}$ and $(c_1+c_2)\mathbf{a} = c_1\mathbf{a}+c_2\mathbf{a}$ for all $\mathbf{a}, \mathbf{b} \in V$ and $c, c_1, c_2 \in \mathbb{F}_q$. There is an associative law of the form $(c_1c_2)\mathbf{a} = c_1(c_2\mathbf{a})$ for all $\mathbf{a} \in V$ and $c_1, c_2 \in \mathbb{F}_q$, and finally we must have $1\mathbf{a} = \mathbf{a}$ for the multiplicative identity $1 \in \mathbb{F}_q$ and all $\mathbf{a} \in V$.

Definition 3.2.1 Let V be a vector space over $\mathbb{F}_q$. The vectors $\mathbf{b}_1, \dots, \mathbf{b}_k \in V$ are *linearly independent* over $\mathbb{F}_q$ if

$$c_1\mathbf{b}_1 + \cdots + c_k\mathbf{b}_k = \mathbf{0}$$

with $c_1, \dots, c_k \in \mathbb{F}_q$ implies that $c_i = 0$ for $1 \leq i \leq k$. The vectors $\mathbf{b}_1, \dots, \mathbf{b}_k \in V$ are *linearly dependent* over $\mathbb{F}_q$ if they are not linearly independent over $\mathbb{F}_q$.

Definition 3.2.2 Let V be a vector space over $\mathbb{F}_q$. The vectors $\mathbf{a}_1, \dots, \mathbf{a}_m \in V$ *generate* V if for every $\mathbf{v} \in V$ there exist $c_1, \dots, c_m \in \mathbb{F}_q$ such that

$$c_1\mathbf{a}_1 + \cdots + c_m\mathbf{a}_m = \mathbf{v}. \tag{3.2}$$

If there exist vectors $\mathbf{a}_1, \dots, \mathbf{a}_m \in V$ that generate V, then V is called *finite-dimensional*.

The identity (3.2) is often expressed by saying that $\mathbf{v}$ is a *linear combination* (over $\mathbb{F}_q$) of $\mathbf{a}_1, \dots, \mathbf{a}_m$. Suppose now that $\mathbf{a}_1, \dots, \mathbf{a}_m$ generate $V \neq \{\mathbf{0}\}$. Let $\{\mathbf{b}_1, \dots, \mathbf{b}_k\}$ be a subset of $\{\mathbf{a}_1, \dots, \mathbf{a}_m\}$ of minimal size such that $\mathbf{b}_1, \dots, \mathbf{b}_k$ generate V. Then $\mathbf{b}_1, \dots, \mathbf{b}_k$ are linearly independent over $\mathbb{F}_q$, for if we had

$$c_1\mathbf{b}_1 + \cdots + c_k\mathbf{b}_k = \mathbf{0}$$

with $c_1, \dots, c_k \in \mathbb{F}_q$ and some $c_i \neq 0$, then either $k = 1$ and $\mathbf{b}_1 = \mathbf{0}$, a contradiction to $V \neq \{\mathbf{0}\}$, or otherwise $k \geq 2$ and $\mathbf{b}_i$ is a linear combination over $\mathbb{F}_q$ of $\mathbf{b}_1, \dots, \mathbf{b}_{i-1}, \mathbf{b}_{i+1}, \dots, \mathbf{b}_k$, which implies that the latter set of $k-1$ vectors already generates V, a contradiction to the minimality condition above. Thus, for every finite-dimensional vector space $V \neq \{\mathbf{0}\}$ there exists a set of vectors with the properties enunciated in the following definition.

Definition 3.2.3 The vectors $\mathbf{b}_1, \dots, \mathbf{b}_k \in V$ form a *basis* of the finite-dimensional vector space $V \neq \{\mathbf{0}\}$ over $\mathbb{F}_q$ if $\mathbf{b}_1, \dots, \mathbf{b}_k$ are linearly independent over $\mathbb{F}_q$ and generate V.

Proposition 3.2.4 *The number of vectors in a basis of a finite-dimensional vector space $V \neq \{\mathbf{0}\}$ over $\mathbb{F}_q$ depends only on V, or in other words, any two bases of V contain the same number of vectors.*

Proof. Suppose that $\mathbf{b}_1, \dots, \mathbf{b}_k \in V$ form a basis of V. Then every $\mathbf{v} \in V$ can be written as a linear combination

$$c_1\mathbf{b}_1 + \cdots + c_k\mathbf{b}_k = \mathbf{v} \tag{3.3}$$

with $c_1, \dots, c_k \in \mathbb{F}_q$. We claim that this representation is unique. Assume we have also

$$c_1'\mathbf{b}_1 + \cdots + c_k'\mathbf{b}_k = \mathbf{v}$$

with $c_1', \dots, c_k' \in \mathbb{F}_q$. By subtracting these two identities, we obtain

$$(c_1 - c_1')\mathbf{b}_1 + \cdots + (c_k - c_k')\mathbf{b}_k = \mathbf{0}.$$

Since $\mathbf{b}_1, \dots, \mathbf{b}_k$ are linearly independent over $\mathbb{F}_q$, it follows from Definition 3.2.1 that $c_i - c_i' = 0$ for $1 \leq i \leq k$, that is, $c_i = c_i'$ for $1 \leq i \leq k$. Thus, the claim concerning the unique representation in (3.3) is established. Consequently, the number of vectors in V is equal to the number of k-tuples $(c_1, \dots, c_k)$ of elements of $\mathbb{F}_q$, and so it is equal to q^k. It remains to observe that k is uniquely determined by V. □

Definition 3.2.5 The *dimension* $\dim(V)$ of a finite-dimensional vector space $V \neq \{\mathbf{0}\}$ over $\mathbb{F}_q$ is the number of vectors in any basis of V. The dimension $\dim(V)$ of $V = \{\mathbf{0}\}$ is defined to be 0.

Proposition 3.2.6 *The number of vectors in a finite-dimensional vector space V over $\mathbb{F}_q$ is equal to $q^{\dim(V)}$.*

Proof. This is trivial for $\dim(V) = 0$, and for $\dim(V) \geq 1$ it was shown in the proof of Proposition 3.2.4. □

Let $V \neq \{\mathbf{0}\}$ be a k-dimensional vector space over $\mathbb{F}_q$ and let $\mathbf{b}_1, \ldots, \mathbf{b}_k \in V$ form a basis of V. If we fix the order of the basis vectors, then we speak of the *ordered basis* $\{\mathbf{b}_1, \ldots, \mathbf{b}_k\}$ of V over $\mathbb{F}_q$. We have shown in the proof of Proposition 3.2.4 that there is a one-to-one correspondence provided by (3.3) between the vectors $\mathbf{v} \in V$ and the vectors $(c_1, \ldots, c_k) \in \mathbb{F}_q^k$. The vector $(c_1, \ldots, c_k)$ is called the *coordinate vector* of $\mathbf{v}$ relative to the ordered basis $\{\mathbf{b}_1, \ldots, \mathbf{b}_k\}$.

Remark 3.2.7 For a positive integer k, the finite field $\mathbb{F}_q$ is a subfield of $\mathbb{F}_{q^k}$ (see Proposition 1.4.31). It is of interest to observe that $\mathbb{F}_{q^k}$ can be viewed as a vector space over $\mathbb{F}_q$, by letting the addition of vectors be the addition in $\mathbb{F}_{q^k}$ and by letting the multiplication of a vector by a scalar from $\mathbb{F}_q$ be the multiplication in $\mathbb{F}_{q^k}$. It is straightforward to check that all properties of a vector space are satisfied. According to Proposition 3.2.6, the dimension of the vector space $\mathbb{F}_{q^k}$ over $\mathbb{F}_q$ is k. A basis of $\mathbb{F}_{q^k}$ can be obtained as follows. Let $f(x) \in \mathbb{F}_q[x]$ be an irreducible polynomial over $\mathbb{F}_q$ of degree k. Then $\mathbb{F}_{q^k}$ can be identified with the residue class field $\mathbb{F}_q[x]/(f(x))$ (see Remark 1.4.46). Thus, the elements of $\mathbb{F}_{q^k}$ can be taken to be the polynomials $\sum_{j=0}^{k-1} a_j x^j$, with all $a_j \in \mathbb{F}_q$, in the least residue system modulo $f(x)$. Given this description of $\mathbb{F}_{q^k}$, it is obvious that $1, x, x^2, \ldots, x^{k-1}$ form a basis of $\mathbb{F}_{q^k}$. Since $f(x) = 0$ in $\mathbb{F}_q[x]/(f(x))$, we can think of x also as a root $\alpha \in \mathbb{F}_{q^k}$ of $f(x)$. Then a basis of $\mathbb{F}_{q^k}$ is formed by $1, \alpha, \alpha^2, \ldots, \alpha^{k-1}$, and $f(x)$ is the minimal polynomial of α over $\mathbb{F}_q$.

Definition 3.2.8 Let V be a vector space over $\mathbb{F}_q$. A subset W of V is a (*linear*) *subspace* of V if W is a vector space over $\mathbb{F}_q$ under the operations inherited from V.

Remark 3.2.9 A subspace of V must always contain the zero vector $\mathbf{0}$ of V. There are two trivial subspaces of V, namely $\{\mathbf{0}\}$ (the *zero subspace*) and V itself. If V is a finite-dimensional vector space over $\mathbb{F}_q$ and W is a subspace of V, then $\dim(W) \leq \dim(V)$ by Proposition 3.2.6.

3.2.2 Fundamental properties of linear codes

After these preparations, let us shift into a higher gear and move on to the important family of linear codes. The basic vector spaces for the theory of linear codes are the Hamming spaces $\mathbb{F}_q^n$, where n is some positive integer. The two operations in the vector space $\mathbb{F}_q^n$ are defined coordinatewise. Thus, if $\mathbf{u} = (u_1, \ldots, u_n) \in \mathbb{F}_q^n$ and $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{F}_q^n$, then the vector addition is defined by

$$\mathbf{u} + \mathbf{v} = (u_1 + v_1, \ldots, u_n + v_n) \in \mathbb{F}_q^n,$$

and the multiplication by a scalar $c \in \mathbb{F}_q$ is defined by

$$c\mathbf{v} = (cv_1, \ldots, cv_n) \in \mathbb{F}_q^n.$$

It is straightforward to verify that, with these operations, $\mathbb{F}_q^n$ satisfies all properties of a vector space over $\mathbb{F}_q$.

Since $\mathbb{F}_q^n$ contains exactly q^n vectors, it follows from Proposition 3.2.6 that $\dim(\mathbb{F}_q^n) = n$. Thus, every basis of $\mathbb{F}_q^n$ consists of n vectors. The *standard basis* of $\mathbb{F}_q^n$ is formed by the vectors $\mathbf{s}_1, \ldots, \mathbf{s}_n \in \mathbb{F}_q^n$, where $\mathbf{s}_j$, $j = 1, \ldots, n$, has jth coordinate equal to 1 and all other coordinates equal to 0.

Definition 3.2.10 A *linear code* C over $\mathbb{F}_q$ is a nonzero subspace of $\mathbb{F}_q^n$ for some positive integer n. The dimension $\dim(C)$ of C as a vector space over $\mathbb{F}_q$ is called the *dimension* of the linear code C.

If $C \subseteq \mathbb{F}_q^n$ is a linear code over $\mathbb{F}_q$ of length n, then it follows from Remark 3.2.9 that the dimension k of C satisfies $1 \leq k \leq n$. It is convenient to call C a *linear* $[n, k]$ *code* over $\mathbb{F}_q$. If we want to point out in addition that the minimum distance of C is d, then we speak of a *linear* $[n, k, d]$ *code* over $\mathbb{F}_q$.

Example 3.2.11 The repetition code $C = \{(a, \ldots, a) \in \mathbb{F}_q^n : a \in \mathbb{F}_q\}$ in Example 3.1.4 has dimension 1 and a basis of C is formed by the single all-one vector $(1, \ldots, 1)$. Therefore we can say that C is a linear $[n, 1]$ code over $\mathbb{F}_q$. By Example 3.1.11 we know that $d(C) = n$, and so C is a linear $[n, 1, n]$ code over $\mathbb{F}_q$.

Example 3.2.12 Consider the binary code C in Example 3.1.16. It is clear that C is a linear code over $\mathbb{F}_2$ with basis $\mathbf{c}_2, \mathbf{c}_3$. Therefore $\dim(C) = 2$, and so C is a linear $[5, 2, 3]$ code over $\mathbb{F}_2$.

Let C be a linear $[n, k]$ code over $\mathbb{F}_q$. Then by Proposition 3.2.6, C contains exactly q^k codewords. In the setting described in Section 3.1, we can then take $\mathbb{F}_q^k$ as the set of messages. Furthermore, it is common practice to let the encoder ψ be a linear transformation from $\mathbb{F}_q^k$ into $\mathbb{F}_q^n$. We recall from linear algebra that if V_1 and V_2 are arbitrary vector spaces over $\mathbb{F}_q$, then a *linear transformation* from V_1 into V_2 is a map $\lambda : V_1 \to V_2$ such that $\lambda(\mathbf{a} + \mathbf{b}) = \lambda(\mathbf{a}) + \lambda(\mathbf{b})$ for all $\mathbf{a}, \mathbf{b} \in V_1$ and $\lambda(c\mathbf{a}) = c\lambda(\mathbf{a})$ for all $c \in \mathbb{F}_q$ and $\mathbf{a} \in V_1$.

If we choose the encoder $\psi : \mathbb{F}_q^k \to \mathbb{F}_q^n$ to be a linear transformation, then encoding (that is, the computation of $\psi(\mathbf{a})$ for $\mathbf{a} \in \mathbb{F}_q^k$) becomes a simple task. Let $\mathbf{s}_1, \ldots, \mathbf{s}_k$ be the standard basis of $\mathbb{F}_q^k$. Then we precompute $\psi(\mathbf{s}_1), \ldots, \psi(\mathbf{s}_k)$. For an arbitrary input $\mathbf{a} = (a_1, \ldots, a_k) \in \mathbb{F}_q^k$ of the encoder, we can write

$$\mathbf{a} = a_1\mathbf{s}_1 + \cdots + a_k\mathbf{s}_k,$$

and so the properties of a linear transformation imply that

$$\psi(\mathbf{a}) = a_1\psi(\mathbf{s}_1) + \cdots + a_k\psi(\mathbf{s}_k). \tag{3.4}$$

We will see a bit later that this computation can also be conveniently described in terms of matrix algebra.

Life is easier with linear codes, and this holds not only for encoding, but also for most other computational tasks for codes. Consider, for instance, the problem of determining the minimum distance of a code. This is greatly facilitated by the following concept and the subsequent theorem valid for all linear codes.

Definition 3.2.13 The *Hamming weight* $w(\mathbf{v})$ of $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{F}_q^n$ is defined to be the number of nonzero coordinates of $\mathbf{v}$, that is,

$$w(\mathbf{v}) = \#\{1 \leq j \leq n : v_j \neq 0\}.$$

By comparing this definition with the definition of the Hamming distance in Definition 3.1.5, we see that if $\mathbf{u}, \mathbf{v} \in \mathbb{F}_q^n$, then

$$w(\mathbf{v}) = d(\mathbf{v}, \mathbf{0}) \quad \text{and} \quad d(\mathbf{u}, \mathbf{v}) = w(\mathbf{u} - \mathbf{v}). \tag{3.5}$$

Theorem 3.2.14 *If C is a linear code over $\mathbb{F}_q$, then its minimum distance $d(C)$ satisfies*

$$d(C) = w(C) := \min_{\mathbf{c} \in C \setminus \{\mathbf{0}\}} w(\mathbf{c}),$$

that is, $d(C)$ is equal to the minimum Hamming weight of a nonzero codeword in C.

Proof. By definition, there exist distinct codewords $\mathbf{c}_1, \mathbf{c}_2 \in C$ such that $d(C) = d(\mathbf{c}_1, \mathbf{c}_2)$. We apply the second identity in (3.5) and get

$$d(C) = d(\mathbf{c}_1, \mathbf{c}_2) = w(\mathbf{c}_1 - \mathbf{c}_2) \geq w(C)$$

since $\mathbf{c}_1 - \mathbf{c}_2 \in C$ for a linear code C. Conversely, there exists a codeword $\mathbf{c} \in C \setminus \{\mathbf{0}\}$ with $w(C) = w(\mathbf{c})$, and then the first identity in (3.5) yields

$$w(C) = w(\mathbf{c}) = d(\mathbf{c}, \mathbf{0}) \geq d(C)$$

since $\mathbf{0} \in C$ for a linear code C. □

Remark 3.2.15 The number $w(C)$ introduced in Theorem 3.2.14 is called the *Hamming weight* of the linear code C. The Hamming weight $w(C)$ is defined in exactly the same way for every nonlinear code C that contains at least one nonzero codeword. The minimum distance and the Hamming weight of a nonlinear code can, for a suitable choice of the code, be as far apart as we desire. For instance, take an arbitrary prime power q and an integer $n \geq 2$ and let $C \subseteq \mathbb{F}_q^n$ be the nonlinear code consisting of the two codewords

$$\mathbf{c}_1 = (1, \ldots, 1) \in \mathbb{F}_q^n, \quad \mathbf{c}_2 = (1, \ldots, 1, 0) \in \mathbb{F}_q^n.$$

Then $d(C) = 1$ and $w(C) = n - 1$.

Suppose that a linear code C for which we want to determine the minimum distance contains exactly s codewords. If we calculate $d(C)$ by its definition in Definition 3.1.9, then we have to compute $(s^2-s)/2$ Hamming distances between all pairs of distinct codewords in C. If we calculate $d(C)$ by Theorem 3.2.14, then we have to compute only the $s-1$ Hamming distances between the nonzero codewords in C and $\mathbf{0} \in C$. Therefore the minimum distance of a linear code is usually calculated by means of Theorem 3.2.14.

Example 3.2.16 The binary linear code C of length 6 is given by its basis

$$\begin{aligned}\mathbf{b}_1 &= (1,0,0,1,1,0),\\ \mathbf{b}_2 &= (0,1,0,1,0,1),\\ \mathbf{b}_3 &= (0,0,1,0,1,1).\end{aligned}$$

We obtain all codewords in C by forming all linear combinations over $\mathbb{F}_2$ of $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$. Besides $\mathbf{0}, \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$, this yields the codewords

$$\begin{aligned}\mathbf{b}_1+\mathbf{b}_2 &= (1,1,0,0,1,1),\\ \mathbf{b}_1+\mathbf{b}_3 &= (1,0,1,1,0,1),\\ \mathbf{b}_2+\mathbf{b}_3 &= (0,1,1,1,1,0),\\ \mathbf{b}_1+\mathbf{b}_2+\mathbf{b}_3 &= (1,1,1,0,0,0).\end{aligned}$$

For the nonzero codewords in C, only the Hamming weights 3 and 4 appear, and so $d(C) = w(C) = 3$ by Theorem 3.2.14. Thus, C is a linear $[6,3,3]$ code over $\mathbb{F}_2$, and C is 1-error-correcting by Theorem 3.1.14.

3.2.3 Matrices over finite fields

Before we begin with matrix algebra and its importance for linear codes, we introduce an operation on $\mathbb{F}_q^n$ which yields an element of $\mathbb{F}_q$ as the output.

Definition 3.2.17 The *dot product* (or *standard inner product*) of $\mathbf{u} = (u_1, \ldots, u_n) \in \mathbb{F}_q^n$ and $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{F}_q^n$ is defined by

$$\mathbf{u} \cdot \mathbf{v} = u_1v_1 + \cdots + u_nv_n \in \mathbb{F}_q.$$

Proposition 3.2.18 *The dot product on* $\mathbb{F}_q^n$ *has the following properties:*
(i) $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$ *for all* $\mathbf{u}, \mathbf{v} \in \mathbb{F}_q^n$*;*
(ii) $\mathbf{u} \cdot (\mathbf{v}_1 + \mathbf{v}_2) = \mathbf{u} \cdot \mathbf{v}_1 + \mathbf{u} \cdot \mathbf{v}_2$ *for all* $\mathbf{u}, \mathbf{v}_1, \mathbf{v}_2 \in \mathbb{F}_q^n$*;*
(iii) $\mathbf{u} \cdot (c\mathbf{v}) = c(\mathbf{u} \cdot \mathbf{v})$ *for all* $\mathbf{u}, \mathbf{v} \in \mathbb{F}_q^n$ *and* $c \in \mathbb{F}_q$.

Proof. All properties are obvious from the definition of the dot product. □

Proposition 3.2.18 implies that the dot product $\mathbf{u} \cdot \mathbf{v}$ is *bilinear*, that is, it is linear in both the first vector $\mathbf{u}$ and the second vector $\mathbf{v}$.

Remark 3.2.19 The dot product of two nonzero vectors can turn out to be 0. For instance, in $\mathbb{F}_3^2$ we have

$$(1,1) \cdot (2,1) = 1 \cdot 2 + 1 \cdot 1 = 0.$$

Two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{F}_q^n$ with $\mathbf{u} \cdot \mathbf{v} = 0$ are said to be *orthogonal*.

Matrix algebra is standard material in a course on linear algebra, but often only real and complex matrices are treated. Here we briefly review matrices over finite fields. A *$k \times n$ matrix* over $\mathbb{F}_q$ is a rectangular array consisting of k rows and n columns, where k and n are positive integers and each entry of the array is an element of $\mathbb{F}_q$. For a $k \times n$ matrix A over $\mathbb{F}_q$, the (i,j) entry of A, that is, the entry in the ith row and jth column of A, is usually denoted by a_{ij}, and the whole matrix can be written as $A = (a_{ij})_{1 \leq i \leq k,\, 1 \leq j \leq n}$.

There are various operations that can be performed for matrices over finite fields, and they are completely analogous to those for real and complex matrices. The $k \times n$ matrix $A = (a_{ij})_{1 \leq i \leq k,\, 1 \leq j \leq n}$ over $\mathbb{F}_q$ is multiplied by a scalar $c \in \mathbb{F}_q$ by multiplying each entry of A by c, that is,

$$cA = (ca_{ij})_{1 \leq i \leq k,\, 1 \leq j \leq n}.$$

Two $k \times n$ matrices $A = (a_{ij})_{1 \leq i \leq k,\, 1 \leq j \leq n}$ and $B = (b_{ij})_{1 \leq i \leq k,\, 1 \leq j \leq n}$ over $\mathbb{F}_q$ are added by adding corresponding entries, that is,

$$A + B = (a_{ij} + b_{ij})_{1 \leq i \leq k,\, 1 \leq j \leq n}.$$

Two matrices A and B over $\mathbb{F}_q$ can be multiplied if they have compatible sizes, that is, if the number of columns of A is equal to the number of rows of B. Accordingly, let A be a $k \times n$ matrix over $\mathbb{F}_q$ and let B be an $n \times m$ matrix over $\mathbb{F}_q$. Then the product AB is a $k \times m$ matrix over $\mathbb{F}_q$, and the (i,j) entry of AB is equal to the dot product $\mathbf{a}_i \cdot \mathbf{b}_j$, where $\mathbf{a}_i$ is the ith row of A, $\mathbf{b}_j$ is the jth column of B, and both are viewed as vectors in $\mathbb{F}_q^n$. As for real and complex matrices, multiplication of matrices over $\mathbb{F}_q$ is associative, but in general not commutative. Matrix multiplication and addition are linked by (left and right) distributive laws. Scalar multiplication of matrices behaves in the expected way when combined with matrix multiplication and addition. For instance, $A(cB) = c(AB)$ for all matrices A and B over $\mathbb{F}_q$ of compatible sizes and for all $c \in \mathbb{F}_q$.

We can turn every matrix A over $\mathbb{F}_q$ on its side by defining its *transpose* $A^\top$. In detail, if A is a $k \times n$ matrix over $\mathbb{F}_q$, then $A^\top$ is the $n \times k$ matrix over $\mathbb{F}_q$ that is obtained by letting the ith row of A (for $1 \leq i \leq k$) become the ith column of $A^\top$, or equivalently by letting the jth column of A (for $1 \leq j \leq n$) become the jth row of $A^\top$.

Proposition 3.2.20 *The transpose of matrices has the following properties:*
(i) $(cA)^\top = cA^\top$ *for every* $c \in \mathbb{F}_q$ *and every matrix* A *over* $\mathbb{F}_q$*;*
(ii) $(A+B)^\top = A^\top + B^\top$ *for all matrices* A *and* B *over* $\mathbb{F}_q$ *of the same size;*
(iii) $(AB)^\top = B^\top A^\top$ *for all matrices* A *and* B *over* $\mathbb{F}_q$ *of compatible sizes;*
(iv) $(A^\top)^\top = A$ *for every matrix* A *over* $\mathbb{F}_q$*.*

Proof. This is a straightforward verification. □

Vectors can be viewed as special cases of matrices. Thus, a row vector from $\mathbb{F}_q^n$ is interpreted as a $1 \times n$ matrix over $\mathbb{F}_q$ and a column vector from $\mathbb{F}_q^n$ is interpreted as an $n \times 1$ matrix over $\mathbb{F}_q$. An operation that occurs frequently in the theory and practice of linear codes is that of multiplication of a matrix and a vector. This operation is a special case of matrix multiplication, and so it is possible only if the sizes are compatible. In detail, if A is a given $k \times n$ matrix over $\mathbb{F}_q$, then we can multiply it from the left by a $1 \times k$ matrix over $\mathbb{F}_q$ and from the right by an $n \times 1$ matrix over $\mathbb{F}_q$. In other words, the product $\mathbf{a}A$ makes sense for a row vector $\mathbf{a} \in \mathbb{F}_q^k$ and the product $A\mathbf{b}$ makes sense for a column vector $\mathbf{b} \in \mathbb{F}_q^n$. In order to pass from row vectors to column vectors, we can use the transpose. It is common practice to let vector symbols like $\mathbf{a}, \mathbf{b}, \ldots$ denote only row vectors, and then column vectors are obtained by forming the transposes $\mathbf{a}^\top, \mathbf{b}^\top, \ldots$. Thus, for the vector-matrix products and the matrix-vector products above, we typically write $\mathbf{a}A$ and $A\mathbf{b}^\top$, respectively. The dot product $\mathbf{u} \cdot \mathbf{v}$ in Definition 3.2.17 can then be expressed also as the product $\mathbf{u}\mathbf{v}^\top$. If we view a vector as a special matrix, then it is consistent to write a vector $(a_1, \ldots, a_k) \in \mathbb{F}_q^k$ in matrix notation $(a_1 \ \ldots \ a_k)$ without commas.

3.2.4 Generator matrix

Supplied with all these tools from linear algebra, we can now return to the theory of linear codes. Let us first reconsider the issue of encoding for linear codes. We have seen earlier that if the encoder ψ is a linear transformation, then for a message $\mathbf{a} = (a_1, \ldots, a_k) \in \mathbb{F}_q^k$, the corresponding codeword $\psi(\mathbf{a}) \in \mathbb{F}_q^n$ is given by the formula in (3.4). Now we set up the $k \times n$ matrix G over $\mathbb{F}_q$ whose row vectors are $\psi(\mathbf{s}_1), \ldots, \psi(\mathbf{s}_k)$ in this order. Then the expression on the right-hand side of (3.4) is the vector-matrix product $\mathbf{a}G$, and so we obtain

$$\psi(\mathbf{a}) = \mathbf{a}G \qquad \text{for all } \mathbf{a} \in \mathbb{F}_q^k. \tag{3.6}$$

Let $C \subseteq \mathbb{F}_q^n$ be the linear code over $\mathbb{F}_q$ corresponding to ψ, that is, C is the image of ψ by definition. Then it is clear from (3.4) that $\psi(\mathbf{s}_1), \ldots, \psi(\mathbf{s}_k)$ generate C. Since ψ is injective, C contains exactly q^k codewords, and so $\dim(C) = k$ by Proposition 3.2.6. Hence $\psi(\mathbf{s}_1), \ldots, \psi(\mathbf{s}_k)$ form a basis of C. This leads to the following terminology for the matrix G in (3.6).

Definition 3.2.21 Let C be a linear $[n, k]$ code over $\mathbb{F}_q$. Then a $k \times n$ matrix over $\mathbb{F}_q$ whose row vectors form a basis of C is called a *generator matrix* of C.

Consequently, efficient encoding for a linear code proceeds by the following simple algorithm.

Algorithm 3.2.22 (Encoding for Linear Codes) Let C be a linear $[n, k]$ code over $\mathbb{F}_q$.
Step 1: choose a basis of C.
Step 2: set up a $k \times n$ generator matrix G of C by writing the basis vectors of C as row vectors of G.
Step 3: the codeword $\mathbf{c} \in C$ corresponding to the message $\mathbf{a} \in \mathbb{F}_q^k$ is computed as $\mathbf{c} = \psi(\mathbf{a}) = \mathbf{a}G$.

Given the linear code C, a generator matrix G of C can be precomputed by Steps 1 and 2 of the algorithm above. If many codewords have to be computed for a concrete data transmission over a noisy channel, then only Step 3 in the algorithm needs to be carried out repeatedly.

Example 3.2.23 A generator matrix G of the linear $[6, 3]$ code over $\mathbb{F}_2$ introduced in Example 3.2.16 is given by

$$G = \begin{pmatrix} 1\,0\,0\,1\,1\,0 \\ 0\,1\,0\,1\,0\,1 \\ 0\,0\,1\,0\,1\,1 \end{pmatrix}.$$

For the message $\mathbf{a} = (1\ 0\ 1) \in \mathbb{F}_2^3$, we use our encoding algorithm to compute the corresponding codeword

$$\mathbf{c} = \psi(\mathbf{a}) = (1\,0\,1) \begin{pmatrix} 1\,0\,0\,1\,1\,0 \\ 0\,1\,0\,1\,0\,1 \\ 0\,0\,1\,0\,1\,1 \end{pmatrix} = (1\,0\,1\,1\,0\,1).$$

A generator matrix of a linear code is usually not unique (this is why we speak of *a* generator matrix and not of *the* generator matrix), since a linear code can have many different bases in general.

Example 3.2.24 It is easily checked that another generator matrix G' of the binary linear code in Examples 3.2.16 and 3.2.23 is given by

$$G' = \begin{pmatrix} 1\,0\,0\,1\,1\,0 \\ 1\,1\,0\,0\,1\,1 \\ 0\,0\,1\,0\,1\,1 \end{pmatrix}.$$

Some linear codes have a generator matrix that is of a special form which turns out to be unique. Before we introduce this special form in the following definition, we recall that an *identity matrix* is a square matrix (that is, the number of rows is equal to the number of columns) for which the entries on the main diagonal (that is, the diagonal running from the upper left corner to the lower right corner) are equal to 1 and all other entries are equal to 0.

Definition 3.2.25 A $k \times n$ generator matrix G over $\mathbb{F}_q$ of the form

$$G = (I_k \mid A)$$

with the $k \times k$ identity matrix I_k over $\mathbb{F}_q$ and some $k \times (n-k)$ matrix A over $\mathbb{F}_q$ is said to be in *standard form.*

For instance, the 3×6 generator matrix G over $\mathbb{F}_2$ in Example 3.2.23 is in standard form. A $k \times n$ generator matrix G in standard form affords a speedup in encoding, since the fact that the first k columns of G come from an identity matrix implies that the word consisting of the first k coordinates of $\psi(\mathbf{a})$ is equal to $\mathbf{a}$. Thus, only $n-k$ coordinates of $\psi(\mathbf{a})$ have to be computed. This feature is illustrated by Example 3.2.23 where the first three coordinates of $\mathbf{c}$ are for free since they are equal to the coordinates of $\mathbf{a}$ in the same order. Unfortunately, not every linear code has a generator matrix in standard form.

Example 3.2.26 Let C be the binary linear $[3,2]$ code with basis $\mathbf{b}_1 = (1,0,0)$, $\mathbf{b}_2 = (1,0,1)$. Then C has six possible generator matrices:

$$\begin{pmatrix} 1\,0\,0 \\ 1\,0\,1 \end{pmatrix}, \begin{pmatrix} 1\,0\,1 \\ 1\,0\,0 \end{pmatrix}, \begin{pmatrix} 0\,0\,1 \\ 1\,0\,0 \end{pmatrix}, \begin{pmatrix} 1\,0\,0 \\ 0\,0\,1 \end{pmatrix}, \begin{pmatrix} 0\,0\,1 \\ 1\,0\,1 \end{pmatrix}, \begin{pmatrix} 1\,0\,1 \\ 0\,0\,1 \end{pmatrix}.$$

None of these generator matrices is in standard form.

Remark 3.2.27 A generator matrix in standard form can always be obtained if we consider linear codes up to a notion of equivalence. Two linear $[n,k]$ codes C and C' over $\mathbb{F}_q$ are called *equivalent* if the codewords in C can be transformed into the codewords in C' by applying a fixed permutation of the coordinates. Then we claim that every linear code C is equivalent to a linear code C' such that C' has a generator matrix in standard form. This is proved by a simple procedure in matrix theory, namely that of transforming a matrix (in this case a generator matrix of C) into reduced echelon form by elementary row operations. Recall that an *elementary row operation* is any one of the following three operations: (i) interchanging two rows; (ii) multiplying a row by a nonzero scalar; (iii) replacing a row by its sum with a scalar multiple of another row. The resulting matrix G in reduced echelon form is still a generator matrix of C. By a suitable permutation of the columns of G (and thus by passing to an equivalent linear code C'), we get a generator matrix G' of C' in standard form.

Example 3.2.28 Consider the linear code C in Example 3.2.26. By interchanging the second and third coordinates of all codewords in C, we get an equivalent linear code C'. If we apply this permutation of the coordinates to the fourth generator matrix in Example 3.2.26, then we obtain the generator matrix

$$G' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

of C' which is in standard form.

Example 3.2.29 Let C be the binary linear $[5, 3]$ code with generator matrix

$$G = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \end{pmatrix}.$$

We transform G into reduced echelon form by elementary row operations. To start with, the second row of G and the third row of G are changed by adding the first row of G to them. This yields

$$G_1 = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

Now the first row of G_1 is changed by adding the sum of the second and third rows of G_1 to it. This yields

$$G_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

Then G_2 is a generator matrix of C in standard form.

In the case where the linear $[n, k]$ code C over $\mathbb{F}_q$ has a generator matrix G in standard form, we know that the encoder $\psi : \mathbb{F}_q^k \to C$ operates by adding $n - k$ coordinates from $\mathbb{F}_q$ to a message $\mathbf{a} \in \mathbb{F}_q^k$; these coordinates depend on G and $\mathbf{a}$. The decoder $\psi^{-1} : C \to \mathbb{F}_q^k$ is then trivial, for all we have to do for a given codeword $\mathbf{c} \in C$ is to delete its last $n - k$ coordinates, and this already yields the corresponding message $\mathbf{a} \in \mathbb{F}_q^k$. For instance, in Example 3.2.23 we delete the last $n - k = 3$ coordinates of $\mathbf{c} = (1\ 0\ 1\ 1\ 0\ 1)$ to obtain $\mathbf{a} = (1\ 0\ 1)$.

3.2.5 The dual code

The dot product introduced in Definition 3.2.17 serves as the basic tool in the duality theory for linear codes.

Definition 3.2.30 The *dual space* $V^\perp$ of a subspace V of $\mathbb{F}_q^n$ is given by

$$V^\perp = \{\mathbf{u} \in \mathbb{F}_q^n : \mathbf{u} \cdot \mathbf{v} = 0 \text{ for all } \mathbf{v} \in V\}.$$

Proposition 3.2.31 *The dual space $V^\perp$ of a subspace V of $\mathbb{F}_q^n$ is again a subspace of $\mathbb{F}_q^n$.*

Proof. It is obvious that $\mathbf{0} \in V^\perp$. Furthermore, it follows from properties of the dot product (see Proposition 3.2.18) that $\mathbf{u} \in V^\perp$ implies $c\mathbf{u} \in V^\perp$ for all $c \in \mathbb{F}_q$, and $\mathbf{u}_1, \mathbf{u}_2 \in V^\perp$ implies $\mathbf{u}_1 + \mathbf{u}_2 \in V^\perp$. □

Example 3.2.32 For $V = \{\mathbf{0}\}$ it is obvious that $V^\perp = \mathbb{F}_q^n$. If $V = \mathbb{F}_q^n$, $\mathbf{u} = (u_1, \dots, u_n) \in V^\perp$, and $\mathbf{s}_1, \dots, \mathbf{s}_n$ is the standard basis of $\mathbb{F}_q^n$, then $u_j = \mathbf{u} \cdot \mathbf{s}_j = 0$ for $1 \le j \le n$, and so $V^\perp = \{\mathbf{0}\}$.

There is a trivial linear code C over $\mathbb{F}_q$ of length n, namely $C = \mathbb{F}_q^n$. It has minimum distance $d(C) = 1$, and so this code is useless since it can neither correct nor detect errors. We emphasize that a linear $[n, k]$ code over $\mathbb{F}_q$ is nontrivial if and only if its dimension k satisfies $1 \le k \le n - 1$.

Definition 3.2.33 If C is a nontrivial linear code over $\mathbb{F}_q$, then the dual space $C^\perp$ of C is called the *dual code* of C.

Theorem 3.2.34 *If C is a nontrivial linear $[n, k]$ code over $\mathbb{F}_q$, then its dual code $C^\perp$ is a linear $[n, n-k]$ code over $\mathbb{F}_q$.*

Proof. Proposition 3.2.31 shows that $C^\perp$ is a subspace of $\mathbb{F}_q^n$. It remains to determine $\dim(C^\perp)$. Since passing to an equivalent linear code does not change $\dim(C)$ and $\dim(C^\perp)$, we can assume that C has a generator matrix in standard form (compare with Remark 3.2.27). Thus, if $\dim(C) = k$, then C has a basis $\mathbf{c}_1, \dots, \mathbf{c}_k$ of the form

$$\mathbf{c}_i = (\mathbf{s}_i, \mathbf{d}_i) \qquad \text{for } 1 \le i \le k$$

with suitable $\mathbf{d}_1, \dots, \mathbf{d}_k \in \mathbb{F}_q^{n-k}$, where $\mathbf{s}_1, \dots, \mathbf{s}_k$ is the standard basis of $\mathbb{F}_q^k$. Consider a fixed $\mathbf{b} \in \mathbb{F}_q^{n-k}$. Let $\mathbf{u} \in \mathbb{F}_q^n$ be of the form

$$\mathbf{u} = (u_1, \dots, u_k, \mathbf{b})$$

for some $u_1, \dots, u_k \in \mathbb{F}_q$. Then $\mathbf{u} \in C^\perp$ if and only if $\mathbf{u} \cdot \mathbf{c}_i = 0$ for $1 \le i \le k$. Because of the special form of the $\mathbf{c}_i$, the latter condition is equivalent to $u_i = -\mathbf{b} \cdot \mathbf{d}_i$ for $1 \le i \le k$. Hence for every fixed $\mathbf{b} \in \mathbb{F}_q^{n-k}$, the coordinates $u_1, \dots, u_k$ of $\mathbf{u} = (u_1, \dots, u_k, \mathbf{b}) \in C^\perp$ are uniquely determined. It follows that $C^\perp$ contains exactly q^{n-k} vectors, and so $\dim(C^\perp) = n - k$ by Proposition 3.2.6. □

Corollary 3.2.35 *Every nontrivial linear code C over $\mathbb{F}_q$ satisfies $(C^\perp)^\perp = C$.*

Proof. From Theorem 3.2.34 we obtain

$$\dim((C^\perp)^\perp) = n - \dim(C^\perp) = n - (n - \dim(C)) = \dim(C).$$

We complete the proof by showing that $C \subseteq (C^\perp)^\perp$. In order to prove that $\mathbf{c} \in C$ implies $\mathbf{c} \in (C^\perp)^\perp$, we have to verify that $\mathbf{c} \cdot \mathbf{u} = 0$ for all $\mathbf{u} \in C^\perp$. But this follows from the definition of $C^\perp$. □

3.2.6 Parity-check matrix

Besides a generator matrix, there is another important type of matrix attached to a linear code (on condition that the linear code is nontrivial), namely a parity-check matrix.

Definition 3.2.36 Let C be a nontrivial linear $[n, k]$ code over $\mathbb{F}_q$. Then an $(n - k) \times n$ matrix over $\mathbb{F}_q$ is a *parity-check matrix* of C if it is a generator matrix of the dual code $C^\perp$.

Theorem 3.2.37 *Let H be a parity-check matrix of a nontrivial linear $[n, k]$ code C over $\mathbb{F}_q$ and let $\mathbf{v} \in \mathbb{F}_q^n$. Then $\mathbf{v} \in C$ if and only if $\mathbf{v}H^\top = \mathbf{0}$.*

Proof. Note that, by definition, the row vectors $\mathbf{h}_1, \ldots, \mathbf{h}_{n-k}$ of H form a basis of $C^\perp$. Moreover, $\mathbf{h}_1^\top, \ldots, \mathbf{h}_{n-k}^\top$ are the column vectors of $H^\top$. If $\mathbf{v} \in C$, then $\mathbf{v} \cdot \mathbf{h}_j = 0$ and so $\mathbf{v}\mathbf{h}_j^\top = 0$ for $1 \leq j \leq n - k$. This means that $\mathbf{v}H^\top = \mathbf{0}$. Conversely, if $\mathbf{v}H^\top = \mathbf{0}$, then $\mathbf{v}\mathbf{h}_j^\top = 0$ and so $\mathbf{v} \cdot \mathbf{h}_j = 0$ for $1 \leq j \leq n - k$. Since $\mathbf{h}_1, \ldots, \mathbf{h}_{n-k}$ generate $C^\perp$, we deduce that $\mathbf{v} \cdot \mathbf{u} = 0$ for all $\mathbf{u} \in C^\perp$. This yields $\mathbf{v} \in (C^\perp)^\perp = C$ by Corollary 3.2.35. □

Corollary 3.2.38 *Let C be a nontrivial linear $[n, k]$ code over $\mathbb{F}_q$. Then for every generator matrix G of C and every parity-check matrix H of C, the identity $GH^\top = O_{k\times(n-k)}$ holds, where $O_{k\times(n-k)}$ is the $k \times (n - k)$ zero matrix over $\mathbb{F}_q$.*

Proof. For every row vector $\mathbf{v}$ of G, we obtain $\mathbf{v}H^\top = \mathbf{0}$ by Theorem 3.2.37, and this shows the desired result. □

Remark 3.2.39 By using the property of the transpose stated in Proposition 3.2.20(iii), we see that the condition $\mathbf{v}H^\top = \mathbf{0}$ in Theorem 3.2.37 can be written also as $H\mathbf{v}^\top = \mathbf{0}^\top$. This can be interpreted as saying that the linear code C in Theorem 3.2.37 forms what is called in linear algebra the null space (or the kernel) of the matrix H. In view of the same property of the transpose, the identity $GH^\top = O_{k\times(n-k)}$ in Corollary 3.2.38 is equivalent to $HG^\top = O_{(n-k)\times k}$.

How can we obtain a parity-check matrix of a given nontrivial linear code C explicitly? By turning to an equivalent linear code, we can assume that C has a generator matrix in standard form (compare with Remark 3.2.27). Then a formula for a parity-check matrix of C is provided by the following theorem. We note that once we know a parity-check matrix H of C, we have also an explicit description of the dual code $C^{\perp}$ of C, since $C^{\perp}$ consists of all linear combinations of the row vectors of H.

Theorem 3.2.40 *If C is a nontrivial linear $[n,k]$ code over $\mathbb{F}_q$ with generator matrix $G = (I_k \mid A)$ in standard form, then $H = (-A^{\top} \mid I_{n-k})$ is a parity-check matrix of C.*

Proof. It is clear from the form of H that the $n-k$ row vectors of H are linearly independent over $\mathbb{F}_q$. In order to prove that H is a parity-check matrix of C (or, by definition, that H is a generator matrix of $C^{\perp}$), it remains to verify that each row vector of H is orthogonal to each row vector of G. Now

$$HG^{\top} = (-A^{\top} \mid I_{n-k}) \begin{pmatrix} I_k \\ A^{\top} \end{pmatrix} = -A^{\top} I_k + I_{n-k} A^{\top} = -A^{\top} + A^{\top} = O_{(n-k)\times k},$$

and this implies the desired property. □

Definition 3.2.41 An $(n-k) \times n$ parity-check matrix H over $\mathbb{F}_q$ of the form

$$H = (B \mid I_{n-k})$$

with the $(n-k) \times (n-k)$ identity matrix I_{n-k} over $\mathbb{F}_q$ and some $(n-k) \times k$ matrix B over $\mathbb{F}_q$ is said to be in *standard form*.

Remark 3.2.42 It follows from Remark 3.2.27 and Theorem 3.2.40 that for every nontrivial linear code there exists an equivalent linear code which has a parity-check matrix in standard form.

Example 3.2.43 Consider the binary linear $[5,3]$ code C in Example 3.2.29. We have seen in that example that C has a generator matrix

$$G_2 = \begin{pmatrix} 1\,0\,0\,0\,1 \\ 0\,1\,0\,1\,1 \\ 0\,0\,1\,1\,1 \end{pmatrix}$$

in standard form. Theorem 3.2.40 now yields a parity-check matrix

$$H = \begin{pmatrix} 0\,1\,1\,1\,0 \\ 1\,1\,1\,0\,1 \end{pmatrix}$$

of C in standard form. By forming the four possible linear combinations over $\mathbb{F}_2$ of the row vectors of H, we find out that the list of all codewords in the dual code $C^\perp$ is given by

$$\mathbf{c}_1 = (0\ 0\ 0\ 0\ 0),\ \ \mathbf{c}_2 = (0\ 1\ 1\ 1\ 0),$$
$$\mathbf{c}_3 = (1\ 1\ 1\ 0\ 1),\ \ \mathbf{c}_4 = (1\ 0\ 0\ 1\ 1).$$

A parity-check matrix of a nontrivial linear code comes in handy when determining the minimum distance of a linear code. This application is based on the following two results.

Theorem 3.2.44 *Let C be a nontrivial linear code over $\mathbb{F}_q$ with parity-check matrix H and let $d \geq 2$ be an integer. Then $d(C) \geq d$ if and only if any $d-1$ column vectors of H are linearly independent over $\mathbb{F}_q$.*

Proof. Let $\mathbf{h}_1^\top, \dots, \mathbf{h}_n^\top$ be the column vectors of H. Recall from Theorem 3.2.37 that the codewords $\mathbf{c} = (c_1, \dots, c_n) \in C$ are characterized by the property $\mathbf{c}H^\top = \mathbf{0}$, that is,

$$c_1\mathbf{h}_1 + \cdots + c_n\mathbf{h}_n = \mathbf{0}.$$

Thus, if any $d-1$ column vectors of H are linearly independent over $\mathbb{F}_q$, then there is no $\mathbf{c} \in C \setminus \{\mathbf{0}\}$ of Hamming weight $w(\mathbf{c}) \leq d-1$, and so $d(C) \geq d$ by Theorem 3.2.14. Similarly, if there are $d-1$ column vectors of H that are linearly dependent over $\mathbb{F}_q$, then $w(\mathbf{c}) \leq d-1$ for some nonzero $\mathbf{c} \in C$, and therefore $d(C) \leq d-1$. □

Corollary 3.2.45 *Let C be a nontrivial linear code over $\mathbb{F}_q$ with parity-check matrix H and let $d \geq 2$ be an integer. Then $d(C) = d$ if and only if any $d-1$ column vectors of H are linearly independent over $\mathbb{F}_q$ and there exist d column vectors of H that are linearly dependent over $\mathbb{F}_q$.*

Proof. This is an immediate consequence of Theorem 3.2.44. □

Example 3.2.46 Let C be the linear $[6, 3]$ code over $\mathbb{F}_2$ with generator matrix

$$G = \begin{pmatrix} 1\ 0\ 0\ 1\ 1\ 0 \\ 0\ 1\ 0\ 0\ 1\ 1 \\ 0\ 0\ 1\ 1\ 0\ 1 \end{pmatrix}$$

in standard form. By Theorem 3.2.40, a parity-check matrix of C is given by

$$H = \begin{pmatrix} 1\ 0\ 1\ 1\ 0\ 0 \\ 1\ 1\ 0\ 0\ 1\ 0 \\ 0\ 1\ 1\ 0\ 0\ 1 \end{pmatrix}.$$

It is easily verified that any two column vectors of H are linearly independent over $\mathbb{F}_2$. On the other hand, the third column of H is the sum of the first and second column of H, and so $d(C) = 3$ by Corollary 3.2.45. Thus, this code is 1-error-correcting.

3.2.7 The syndrome decoding algorithm

We have learned that linear codes allow fast encoders and decoders. Now we look at the remaining computational procedure in a coding scheme with a linear code, namely the decoding algorithm. It will transpire that reasonably efficient decoding algorithms can be designed for linear codes, and so linear codes achieve goal (iii) stated at the end of Subsection 3.1.2.

We study decoding algorithms for linear codes in the framework of nearest neighbor decoding described in Algorithm 3.1.13. Let C be a nontrivial linear code over $\mathbb{F}_q$ of length n. If the codeword $\mathbf{c} \in C$ is sent over the noisy channel and the word $\mathbf{v} \in \mathbb{F}_q^n$ is received, then

$$\mathbf{e} = \mathbf{v} - \mathbf{c} \tag{3.7}$$

is the *error word* (or *error pattern*). If the received word $\mathbf{v}$ is given, then finding $\mathbf{c}$ is equivalent to finding $\mathbf{e}$. Many decoding algorithms are thus focusing on the error word $\mathbf{e}$. A general property of $\mathbf{e}$ that immediately follows from (3.7) is $\mathbf{e} \in \mathbf{v} + C$, where the latter set is the *coset*

$$\mathbf{v} + C = \{\mathbf{v} + \mathbf{c} : \mathbf{c} \in C\}$$

which can be defined for all $\mathbf{v} \in \mathbb{F}_q^n$. This terminology stems from group theory (compare with Section 1.3). Note that $\mathbb{F}_q^n$, like any vector space, is an abelian group under vector addition. The linear code C is a subgroup of $\mathbb{F}_q^n$. The set $\mathbf{v} + C$ above is exactly the coset (in the sense of group theory) of $\mathbf{v}$ with respect to the subgroup C. As in the general theory of abelian groups, cosets have the following properties which we prove in detail in the present context for your convenience.

Proposition 3.2.47 *Let C be a nontrivial linear code over $\mathbb{F}_q$ of length n. Then:*
(i) *two cosets of C are either identical or they have empty intersection;*
(ii) *if $\mathbf{v}, \mathbf{w} \in \mathbb{F}_q^n$, then $\mathbf{v} - \mathbf{w} \in C$ if and only if $\mathbf{v}$ and $\mathbf{w}$ are in the same coset of C.*

Proof. (i) Consider two cosets $\mathbf{v} + C$ and $\mathbf{w} + C$ and suppose that $\mathbf{u} \in (\mathbf{v} + C) \cap (\mathbf{w} + C)$. From $\mathbf{u} \in \mathbf{v} + C$ we deduce that $\mathbf{u} = \mathbf{v} + \mathbf{c}_0$ for some $\mathbf{c}_0 \in C$, and so

$$\mathbf{u} + C = \{\mathbf{u} + \mathbf{c} : \mathbf{c} \in C\} = \{\mathbf{v} + \mathbf{c}_0 + \mathbf{c} : \mathbf{c} \in C\} = \mathbf{v} + C.$$

Similarly, $\mathbf{u} \in \mathbf{w} + C$ implies that $\mathbf{u} + C = \mathbf{w} + C$. Therefore the cosets $\mathbf{v} + C$ and $\mathbf{w} + C$ are identical since they are both equal to $\mathbf{u} + C$.

(ii) If $\mathbf{v} - \mathbf{w} \in C$, then $\mathbf{v} = \mathbf{w} + \mathbf{c}_0$ for some $\mathbf{c}_0 \in C$, and so $\mathbf{v} \in \mathbf{w} + C$. Also $\mathbf{w} \in \mathbf{w} + C$ since $\mathbf{0} \in C$, and so $\mathbf{v}$ and $\mathbf{w}$ are in the same coset of C. Conversely, if $\mathbf{v}$ and $\mathbf{w}$ are in the same coset $\mathbf{u} + C$ for some $\mathbf{u} \in \mathbb{F}_q^n$, then $\mathbf{v} - \mathbf{u} \in C$ and $\mathbf{w} - \mathbf{u} \in C$, hence $\mathbf{v} - \mathbf{w} = (\mathbf{v} - \mathbf{u}) - (\mathbf{w} - \mathbf{u}) \in C$. □

Let C be a nontrivial linear $[n, k]$ code over $\mathbb{F}_q$. Then Proposition 3.2.47(i) implies that the different cosets of C form a partition of $\mathbb{F}_q^n$. Since each coset of C contains exactly q^k vectors, it follows that there are exactly q^{n-k} different cosets of C.

Let $\mathbf{v} \in \mathbb{F}_q^n$ again be the received word. We have realized that the unknown error word $\mathbf{e}$ belongs to the coset $\mathbf{v} + C$. In conformity with the philosophy of nearest neighbor decoding, we assume that few rather than many errors have occurred in the transmission over the noisy channel. Concretely, we suppose that $\mathbf{e}$ has the smallest Hamming weight within the coset $\mathbf{v} + C$. This leads to the following concept.

Definition 3.2.48 Let C be a nontrivial linear code. A word of minimum Hamming weight within a coset $\mathbf{v}+C$ is called a *coset leader* of $\mathbf{v}+C$. If several words in $\mathbf{v}+C$ have minimum Hamming weight within $\mathbf{v}+C$, we choose one of them arbitrarily as coset leader.

At this stage, we already have a preliminary version of a decoding algorithm for a nontrivial linear code C over $\mathbb{F}_q$ of length n. For a received word $\mathbf{v} \in \mathbb{F}_q^n$, we consider the corresponding coset $\mathbf{v} + C$. The coset leader $\mathbf{e}'$ of $\mathbf{v} + C$ is a most likely error word, and a most likely sent codeword $\mathbf{c}'$ is obtained from (3.7) as $\mathbf{c}' = \mathbf{v} - \mathbf{e}'$.

Consequently, the crucial objects in this context are the coset leaders. Given a nontrivial linear code $C \subseteq \mathbb{F}_q^n$, the coset leaders can in principle be precomputed by inspecting all cosets of C and picking from each coset a word of minimum Hamming weight. This yields the list of all coset leaders. The remaining issue is to figure out the coset leader $\mathbf{e}'$ to which a given $\mathbf{v} \in \mathbb{F}_q^n$ belongs. Note that $\mathbf{e}'$ and $\mathbf{v}$ are in the same coset $\mathbf{v} + C$, and so it is a matter of finding an efficient way of deciding which word from a given list of words (the list of all coset leaders) belongs to a given coset of C (the coset $\mathbf{v} + C$ determined by the received word $\mathbf{v}$). This is achieved by the following notion.

Definition 3.2.49 Let C be a nontrivial linear $[n, k]$ code over $\mathbb{F}_q$ with parity-check matrix H. Then the word $S(\mathbf{v}) = \mathbf{v}H^\top \in \mathbb{F}_q^{n-k}$ is called the *syndrome* of $\mathbf{v} \in \mathbb{F}_q^n$.

Remark 3.2.50 Strictly speaking, since the syndrome depends on the choice of the parity-check matrix H, it would be more precise to denote the syndrome of $\mathbf{v}$ by $S_H(\mathbf{v})$ to signalize this dependence. However, for simplicity of notation, the subscript H is dropped as we tacitly assume that H is fixed in the decoding algorithm for a given nontrivial linear code.

Proposition 3.2.51 *Let C be a nontrivial linear $[n, k]$ code over $\mathbb{F}_q$. Then the syndromes of $\mathbf{v}, \mathbf{w} \in \mathbb{F}_q^n$ satisfy:*
(i) $S(\mathbf{v}) = \mathbf{0}$ *if and only if* $\mathbf{v} \in C$;
(ii) $S(\mathbf{v}) = S(\mathbf{w})$ *if and only if* $\mathbf{v}$ *and* $\mathbf{w}$ *are in the same coset of* C.

Proof. (i) This follows from Theorem 3.2.37.

(ii) By Proposition 3.2.47(ii), $\mathbf{v}$ and $\mathbf{w}$ are in the same coset of C if and only if $\mathbf{v}-\mathbf{w} \in C$. The latter condition is equivalent to $S(\mathbf{v}-\mathbf{w}) = \mathbf{0}$ by part (i), and this is the same as saying that $S(\mathbf{v}) = S(\mathbf{w})$. □

It follows from Proposition 3.2.51 that there is a one-to-one correspondence between the different cosets of C and the different syndromes of words from $\mathbb{F}_q^n$. Each coset of C can thus be uniquely identified with the syndrome of its coset leader. This principle is used in the following refined version of the preliminary decoding algorithm described earlier.

Algorithm 3.2.52 (Syndrome Decoding Algorithm for Linear Codes) Let C be a nontrivial linear code over $\mathbb{F}_q$ of length n and assume that a parity-check matrix of C is known.

Precomputation: compute all coset leaders and the syndrome of each coset leader.

Step 1: for a received word $\mathbf{v} \in \mathbb{F}_q^n$, compute the syndrome $S(\mathbf{v})$.

Step 2: in the list of syndromes of coset leaders, find the coset leader $\mathbf{e}'$ with $S(\mathbf{e}') = S(\mathbf{v})$; then $\mathbf{e}'$ is a most likely error word.

Step 3: compute a most likely sent codeword $\mathbf{c}'$ as $\mathbf{c}' = \mathbf{v} - \mathbf{e}'$.

Example 3.2.53 Let C be the linear $[7,4]$ code over $\mathbb{F}_2$ with generator matrix

$$G = \begin{pmatrix} 1&0&0&0&0&1&1\\ 0&1&0&0&1&0&1\\ 0&0&1&0&1&1&1\\ 0&0&0&1&1&1&0 \end{pmatrix}$$

in standard form. Then Theorem 3.2.40 yields a parity-check matrix

$$H = \begin{pmatrix} 0&1&1&1&1&0&0\\ 1&0&1&1&0&1&0\\ 1&1&1&0&0&0&1 \end{pmatrix}$$

of C. There are exactly $2^{7-4} = 8$ different cosets of C. Next we determine the coset leaders. It turns out that, for this code C, each coset of C has a unique coset leader. The following table lists the coset leaders and their syndromes.

coset leader	syndrome
(0 0 0 0 0 0 0)	(0 0 0)
(1 0 0 0 0 0 0)	(0 1 1)
(0 1 0 0 0 0 0)	(1 0 1)
(0 0 1 0 0 0 0)	(1 1 1)
(0 0 0 1 0 0 0)	(1 1 0)
(0 0 0 0 1 0 0)	(1 0 0)
(0 0 0 0 0 1 0)	(0 1 0)
(0 0 0 0 0 0 1)	(0 0 1)

Suppose that the received word is $\mathbf{v} = (0\ 1\ 1\ 0\ 1\ 1\ 0) \in \mathbb{F}_2^7$. Its syndrome is

$$S(\mathbf{v}) = \mathbf{v}H^\top = (0\ 1\ 1\ 0\ 1\ 1\ 0) \begin{pmatrix} 0\ 1\ 1 \\ 1\ 0\ 1 \\ 1\ 1\ 1 \\ 1\ 1\ 0 \\ 1\ 0\ 0 \\ 0\ 1\ 0 \\ 0\ 0\ 1 \end{pmatrix} = (1\ 0\ 0).$$

This syndrome agrees with the syndrome of the coset leader $\mathbf{e}' = (0\ 0\ 0\ 0\ 1\ 0\ 0)$. Hence $\mathbf{e}'$ is a most likely error word, and a most likely sent codeword is

$$\mathbf{c}' = \mathbf{v} - \mathbf{e}' = (0\ 1\ 1\ 0\ 1\ 1\ 0) - (0\ 0\ 0\ 0\ 1\ 0\ 0) = (0\ 1\ 1\ 0\ 0\ 1\ 0).$$

We can check that $\mathbf{c}'$ is indeed a codeword in C as it is the sum of the second row and the third row of G. By inspecting the parity-check matrix H and using Corollary 3.2.45, we see that $d(C) = 3$, and so C is a 1-error-correcting code. If we assume that the noisy channel allows at most one transmission error in a word of length 7, then we can conclude that $\mathbf{c}'$ is in fact the correct codeword that was sent and that a transmission error occurred in the fifth coordinate of $\mathbf{c}'$. Since G is in standard form, the original message $\mathbf{a} \in \mathbb{F}_2^4$ is then obtained by deleting the last three coordinates of $\mathbf{c}'$, that is, $\mathbf{a} = (0\ 1\ 1\ 0)$.

3.2.8 The MacWilliams identity

We now return to the subject of dual codes (see Definition 3.2.33) which offers many fascinating aspects. We know that a crucial parameter of a code is its minimum distance, which in the case of a linear code is equal to the minimum Hamming weight of a nonzero codeword (see Theorem 3.2.14). It is of interest to determine not only this minimum Hamming weight, but also the complete weight distribution of the linear code. This information is captured by the following notion.

Definition 3.2.54 Let C be a linear code of length n. Then the *weight enumerator* of C is the polynomial

$$A(x) = \sum_{j=0}^{n} A_j x^j \in \mathbb{Z}[x]$$

over the ring $\mathbb{Z}$ of integers, where A_j for $0 \leq j \leq n$ is the number of codewords in C of Hamming weight j.

There is a famous identity that links the weight enumerator of a nontrivial linear code and the weight enumerator of its dual code. This identity was proved by Jessie MacWilliams, one of several prominent female coding theorists, in her Ph.D. thesis [107] and this is no mean achievement for a graduate student.

Theorem 3.2.55 (MacWilliams Identity) *If C is a nontrivial linear $[n, k]$ code over $\mathbb{F}_q$ with weight enumerator $A(x)$, then the weight enumerator $A^{\perp}(x)$ of the dual code $C^{\perp}$ is given by*

$$A^{\perp}(x) = q^{-k}(1+(q-1)x)^n A\left(\frac{1-x}{1+(q-1)x}\right).$$

Proof. Since the dual code $C^{\perp}$ is defined in terms of the dot product on $\mathbb{F}_q^n$, it is not surprising that the proof employs properties of the dot product. Fix a nontrivial additive character χ of $\mathbb{F}_q$. For $\mathbf{u} \in C$ we introduce the polynomial $g_{\mathbf{u}}(x)$ over the field of complex numbers given by

$$g_{\mathbf{u}}(x) = \sum_{\mathbf{v} \in \mathbb{F}_q^n} \chi(\mathbf{u} \cdot \mathbf{v}) x^{w(\mathbf{v})},$$

where $w(\mathbf{v})$ is the Hamming weight of $\mathbf{v} \in \mathbb{F}_q^n$. Then

$$\sum_{\mathbf{u} \in C} g_{\mathbf{u}}(x) = \sum_{\mathbf{u} \in C} \sum_{\mathbf{v} \in \mathbb{F}_q^n} \chi(\mathbf{u} \cdot \mathbf{v}) x^{w(\mathbf{v})} = \sum_{\mathbf{v} \in \mathbb{F}_q^n} x^{w(\mathbf{v})} \sum_{\mathbf{u} \in C} \chi(\mathbf{u} \cdot \mathbf{v}).$$

Consider the inner sum in the last expression. For fixed $\mathbf{v} \in \mathbb{F}_q^n$, the map $\sigma : \mathbf{u} \in C \mapsto \chi(\mathbf{u} \cdot \mathbf{v})$ is a character of the finite abelian group C. If $\mathbf{v} \in C^{\perp}$, then the inner sum is q^k. If $\mathbf{v} \notin C^{\perp}$, then σ is a nontrivial character, and so the inner sum is 0 by the orthogonality relation (1.9). Therefore

$$\sum_{\mathbf{u} \in C} g_{\mathbf{u}}(x) = q^k \sum_{\mathbf{v} \in C^{\perp}} x^{w(\mathbf{v})} = q^k A^{\perp}(x). \tag{3.8}$$

Now we compute the left-hand side of (3.8) in a different way. For $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{F}_q^n$, we write

$$w(\mathbf{v}) = w_1(v_1) + \cdots + w_1(v_n)$$

in analogy with (3.1), where w_1 is the Hamming weight for words of length 1. Then for $\mathbf{u} = (u_1, \ldots, u_n) \in C$ we infer from the definition of $g_{\mathbf{u}}(x)$ that

$$\begin{aligned} g_{\mathbf{u}}(x) &= \sum_{v_1,\ldots,v_n \in \mathbb{F}_q} \chi(u_1v_1 + \cdots + u_nv_n)x^{w_1(v_1)+\cdots+w_1(v_n)} \\ &= \sum_{v_1,\ldots,v_n \in \mathbb{F}_q} \chi(u_1v_1)x^{w_1(v_1)} \cdots \chi(u_nv_n)x^{w_1(v_n)} \\ &= \prod_{j=1}^{n} \big(\sum_{v \in \mathbb{F}_q} \chi(u_jv)x^{w_1(v)} \big). \end{aligned}$$

In the last expression, the inner sum is equal to $1+(q-1)x$ if $u_j = 0$, and for $u_j \neq 0$ it is equal to

$$1 + \big(\sum_{a \in \mathbb{F}_q^*} \chi(a) \big) x = 1 + \big(\sum_{a \in \mathbb{F}_q} \chi(a) - 1 \big) x = 1 - x,$$

again by the orthogonality relation (1.9). Therefore

$$g_{\mathbf{u}}(x) = (1 + (q-1)x)^{n-w(\mathbf{u})}(1-x)^{w(\mathbf{u})} = (1+(q-1)x)^n \left(\frac{1-x}{1+(q-1)x} \right)^{w(\mathbf{u})}.$$

It follows that

$$\begin{aligned} \sum_{\mathbf{u} \in C} g_{\mathbf{u}}(x) &= (1+(q-1)x)^n \sum_{\mathbf{u} \in C} \left(\frac{1-x}{1+(q-1)x} \right)^{w(\mathbf{u})} \\ &= (1+(q-1)x)^n A \left(\frac{1-x}{1+(q-1)x} \right). \end{aligned}$$

By invoking (3.8), we arrive at the desired formula for $A^{\perp}(x)$. □

Example 3.2.56 Let C be the binary linear $[6, 3, 3]$ code in Example 3.2.16. From the complete list of codewords in C given in Example 3.2.16, we see that $A_0 = 1$, $A_1 = A_2 = A_5 = A_6 = 0$, $A_3 = 4$, and $A_4 = 3$ in the notation of Definition 3.2.54. Therefore the weight enumerator of C is the polynomial

$$A(x) = 1 + 4x^3 + 3x^4 \in \mathbb{Z}[x].$$

The MacWilliams identity in Theorem 3.2.55 shows that

$$\begin{aligned} A^{\perp}(x) &= 2^{-3}(1+x)^6 A\left(\frac{1-x}{1+x} \right) \\ &= \frac{1}{8}(1+x)^6 \left[1 + 4\left(\frac{1-x}{1+x} \right)^3 + 3\left(\frac{1-x}{1+x} \right)^4 \right] \\ &= \frac{1}{8} \left[(1+x)^6 + 4(1-x)^3(1+x)^3 + 3(1-x)^4(1+x)^2 \right]. \end{aligned}$$

A straightforward algebraic manipulation yields

$$A^{\perp}(x) = 1 + 4x^3 + 3x^4 \in \mathbb{Z}[x]$$

as the weight enumerator of the dual code $C^{\perp}$. This is an example where C and $C^{\perp}$ have the same weight enumerator and thus the same weight distribution, although the two linear codes C and $C^{\perp}$ are different. For instance, the vector

$$(0, 0, 0, 1, 1, 1) \in \mathbb{F}_2^6$$

belongs to $C^{\perp}$, but not to C.

3.2.9 Self-orthogonal and self-dual codes

We briefly consider nontrivial linear codes for which any two codewords are orthogonal to each other, or where we have the even stronger property that the nontrivial linear code is equal to its dual code.

Definition 3.2.57 A nontrivial linear code C over $\mathbb{F}_q$ is *self-orthogonal* if $C \subseteq C^{\perp}$ and it is *self-dual* if $C = C^{\perp}$.

Proposition 3.2.58 *Let C be a nontrivial linear $[n, k]$ code over $\mathbb{F}_q$. If C is self-orthogonal, then $k \le n/2$, and if C is self-dual, then $k = n/2$. In particular, if C is self-dual, then its length n must be even.*

Proof. If C is self-orthogonal, then $C \subseteq C^{\perp}$ by definition, and a comparison of dimensions yields $k \le n - k$ by Theorem 3.2.34. This implies $k \le n/2$. If C is self-dual, then an analogous argument yields $k = n - k$, and so $k = n/2$. □

Example 3.2.59 A simple example of a binary self-dual code is given by the linear code of length 4 with basis vectors $(1, 0, 1, 0)$ and $(0, 1, 0, 1)$. A simple example of a ternary self-dual code is given by the linear code of length 4 with basis vectors $(1, 0, 1, 1)$ and $(0, 1, 1, 2)$.

Example 3.2.60 Let q be a prime power with $q \equiv 1 \pmod 4$ and let k be a positive integer. We construct a self-dual code C over $\mathbb{F}_q$ of length $2k$ as follows. Choose an element $a \in \mathbb{F}_q^*$ and a primitive element g of $\mathbb{F}_q$ (see Definition 1.4.34). Put $c = g^{(q-1)/4}$. The linear code C is given by its basis $\{\mathbf{b}_1, \dots, \mathbf{b}_k\}$. Here for $1 \le i \le k$, we put

$$\mathbf{b}_i = (0, \dots, 0, a, ca, 0, \dots, 0) \in \mathbb{F}_q^{2k},$$

where the entry a is in position $2i - 1$ and the entry ca is in position $2i$. It is clear that $\mathbf{b}_1, \dots, \mathbf{b}_k$ are linearly independent over $\mathbb{F}_q$, and so $\dim(C) = k$. For

$1 \le i < j \le k$, it is immediately seen that $\mathbf{b}_i \cdot \mathbf{b}_j = 0$. Furthermore, for $1 \le i \le k$ we have

$$\mathbf{b}_i \cdot \mathbf{b}_i = a^2 + c^2 a^2 = 0$$

since $c^2 = g^{(q-1)/2} = -1$. The bilinearity of the dot product implies that $\mathbf{c} \cdot \mathbf{d} = 0$ for all $\mathbf{c}, \mathbf{d} \in C$, and so $C \subseteq C^{\perp}$. A comparison of dimensions shows that $C = C^{\perp}$, hence C is indeed self-dual.

Example 3.2.61 It is easy to construct self-orthogonal codes that are not self-dual, especially if the dimension of the code is low. For instance, take any nonzero vector $\mathbf{b} \in \mathbb{F}_q^n$ with $\mathbf{b} \cdot \mathbf{b} = 0$ and let C be the one-dimensional linear code with basis vector $\mathbf{b}$. Then C is self-orthogonal, but for $n \ge 3$ this code is not self-dual.

Further examples of self-orthogonal codes will be presented in Theorems 3.5.18 and 3.5.19. For further examples of self-dual (and thus self-orthogonal) codes, we refer to Example 3.5.4, Example 3.5.6, Proposition 3.5.21, and Theorem 3.5.27.

3.3 Cyclic codes

3.3.1 Cyclic codes and ideals

It is a plausible principle that the more structure we have for a family of codes, the nicer a theory we can develop for it. In this section, we consider linear codes that have the additional property of being invariant under cyclic shifts. This is why they are called cyclic codes, but maybe they are named also after Shannon's unicycle (see Subsection 3.1.1). The rich theory of cyclic codes involves a fascinating interplay with polynomials over finite fields. First we introduce a convenient notation for cyclic shifts of vectors from $\mathbb{F}_q^n$.

Definition 3.3.1 For every $\mathbf{v} = (v_0, v_1, \dots, v_{n-1}) \in \mathbb{F}_q^n$ and every integer t with $0 \le t \le n-1$, the *cyclic shift* $\mathbf{v}^t$ by t positions is defined by

$$\mathbf{v}^t = (v_{n-t}, v_{n-t+1}, \dots, v_{n-1}, v_0, v_1, \dots, v_{n-t-1}) \in \mathbb{F}_q^n.$$

Note that we have taken it for granted that the cyclic shift is by t positions to the *right*. Cyclic shifts to the *left* are also covered by this definition: $\mathbf{v}^{n-1}$ is the cyclic shift by one position to the left, $\mathbf{v}^{n-2}$ is the cyclic shift by two positions to the left, and so on. Note also that $\mathbf{v}^0 = \mathbf{v}$ for all $\mathbf{v} \in \mathbb{F}_q^n$. Formally, we may put also $\mathbf{v}^n = \mathbf{v}^0$, $\mathbf{v}^{n+1} = \mathbf{v}^1$, and so on.

Definition 3.3.2 A linear code $C \subseteq \mathbb{F}_q^n$ is *cyclic* if $\mathbf{c} \in C$ implies $\mathbf{c}^t \in C$ for $1 \le t \le n-1$.

Remark 3.3.3 For $n \geq 2$ it suffices to request in Definition 3.3.2 that $\mathbf{c} \in C$ implies $\mathbf{c}^1 \in C$, for then $\mathbf{c}^2 = (\mathbf{c}^1)^1 \in C$, $\mathbf{c}^3 = (\mathbf{c}^2)^1 \in C$, and in general $\mathbf{c}^t \in C$ for $1 \leq t \leq n-1$. A linear code over $\mathbb{F}_q$ of length 1 (the only such code is in fact $\mathbb{F}_q$ itself) is automatically cyclic since the condition in Definition 3.3.2 is then vacuously satisfied.

Example 3.3.4 The following are easy examples of cyclic codes:
(i) the binary linear code of length 4 given by

$$\{(0,0,0,0),\ (1,0,1,0),\ (0,1,0,1),\ (1,1,1,1)\};$$

(ii) every repetition code (see Example 3.1.4);
(iii) the trivial linear code $\mathbb{F}_q^n$.

A basic device for the analysis of cyclic codes is a correspondence between vectors from $\mathbb{F}_q^n$ and polynomials from $\mathbb{F}_q[x]_{<n}$, the set of polynomials over $\mathbb{F}_q$ of degree less than n. This correspondence is furnished by the map $\pi : \mathbb{F}_q^n \to \mathbb{F}_q[x]_{<n}$ which is defined by

$$\pi(\mathbf{v}) = \sum_{j=0}^{n-1} v_j x^j \quad \text{for all } \mathbf{v} = (v_0, v_1, \ldots, v_{n-1}) \in \mathbb{F}_q^n. \tag{3.9}$$

Note that $\mathbb{F}_q[x]_{<n}$ is a vector space over $\mathbb{F}_q$, with the vector addition and the multiplication by scalars given by the ordinary addition of polynomials and the multiplication of polynomials by elements of $\mathbb{F}_q$, respectively. Then $\pi : \mathbb{F}_q^n \to \mathbb{F}_q[x]_{<n}$ is a bijective linear transformation, and so in the language of linear algebra the vector spaces $\mathbb{F}_q^n$ and $\mathbb{F}_q[x]_{<n}$ are *isomorphic*. A linear code $C \subseteq \mathbb{F}_q^n$ is a nonzero subspace of $\mathbb{F}_q^n$, and so $\pi(C)$ is a nonzero subspace of $\mathbb{F}_q[x]_{<n}$ with the same dimension as C.

Example 3.3.5 Let C be the binary cyclic code in Example 3.3.4(i). Then

$$\pi(C) = \{0,\ 1+x^2,\ x+x^3,\ 1+x+x^2+x^3\} \subset \mathbb{F}_2[x]_{<4}.$$

It is clear that $\pi(C)$ is a subspace of $\mathbb{F}_2[x]_{<4}$ of dimension 2, with $1+x^2$ and $x+x^3$ forming a basis.

If a linear code $C \subseteq \mathbb{F}_q^n$ has the additional property of being cyclic, then this property can be captured by endowing $\mathbb{F}_q[x]_{<n}$ with an additional operation of multiplication. In the polynomial ring $\mathbb{F}_q[x]$ we have the ordinary multiplication of polynomials. If two polynomials from $\mathbb{F}_q[x]_{<n}$ are multiplied, this may yield an overflow in the sense that the product has degree $\geq n$, and then the product does not belong to $\mathbb{F}_q[x]_{<n}$. In order to obtain a product that is again in $\mathbb{F}_q[x]_{<n}$, we need to modify the ordinary multiplication of polynomials in $\mathbb{F}_q[x]$. This is accomplished by first computing the ordinary product of polynomials in $\mathbb{F}_q[x]$, dividing it by a fixed polynomial over $\mathbb{F}_q$ of degree n, and then taking the remainder as the modified

product. Note that the remainder is a polynomial over $\mathbb{F}_q$ of degree less than n and hence an element of $\mathbb{F}_q[x]_{<n}$. In the theory of cyclic codes, the fixed polynomial over $\mathbb{F}_q$ of degree n is $x^n - 1$. In the terminology of abstract algebra, we are thus turning $\mathbb{F}_q[x]_{<n}$ into the residue class ring $\mathbb{F}_q[x]/(x^n - 1)$ (see Subsection 1.4.3). Formally, $\mathbb{F}_q[x]/(x^n - 1)$ consists of residue classes modulo $x^n - 1$, but we can identify each residue class modulo $x^n - 1$ with a unique element from the least residue system $\mathbb{F}_q[x]_{<n}$ modulo $x^n - 1$, and this is done in the following. Note that, with this identification, addition in $\mathbb{F}_q[x]/(x^n - 1)$ agrees with ordinary addition of polynomials in $\mathbb{F}_q[x]$ since there is no possibility of overflow with addition. The arithmetic operations in $\mathbb{F}_q[x]/(x^n - 1)$ can be expressed also by means of congruences modulo $x^n - 1$ (see again Subsection 1.4.3).

Example 3.3.6 Let $q = 3$ and $n = 4$, so that we are looking at the residue class ring $\mathbb{F}_3[x]/(x^4 - 1)$ identified with $\mathbb{F}_3[x]_{<4}$. Since $-1 = 2$ in $\mathbb{F}_3$, we consider equivalently the residue class ring $\mathbb{F}_3[x]/(x^4 + 2)$. Let the two elements $f_1(x) = x^2 + x + 1$ and $f_2(x) = x^3 + 2x + 1$ of $\mathbb{F}_3[x]/(x^4 + 2)$ be given. Then addition in $\mathbb{F}_3[x]/(x^4 + 2)$ yields

$$f_1(x) + f_2(x) = x^3 + x^2 + 2,$$

just like for ordinary addition of polynomials in $\mathbb{F}_3[x]$. To multiply $f_1(x)$ and $f_2(x)$ in $\mathbb{F}_3[x]/(x^4 + 2)$, we first compute the ordinary product

$$f_1(x)f_2(x) = (x^2 + x + 1)(x^3 + 2x + 1) = x^5 + x^4 + 1 \in \mathbb{F}_3[x].$$

Here we have an overflow, hence we need to divide $x^5 + x^4 + 1$ by $x^4 + 2$. This division with remainder in $\mathbb{F}_3[x]$ yields

$$x^5 + x^4 + 1 = (x + 1)(x^4 + 2) + x + 2.$$

Therefore in $\mathbb{F}_3[x]/(x^4 + 2)$ we obtain $f_1(x)f_2(x) = x + 2$, the remainder in the division above. As noted before, the arithmetic operations in $\mathbb{F}_3[x]/(x^4 + 2)$ can be expressed also in the language of congruences modulo $x^4 + 2$, so that we can write $f_1(x)f_2(x) \equiv x + 2 \pmod{x^4 + 2}$.

After the identification of $\mathbb{F}_q[x]_{<n}$ with $\mathbb{F}_q[x]/(x^n - 1)$, the map π in (3.9) is now viewed as a map $\pi : \mathbb{F}_q^n \to \mathbb{F}_q[x]/(x^n - 1)$. It is still a bijective linear transformation between these two vector spaces over $\mathbb{F}_q$.

In order to proceed further, we need the concept of an ideal of a commutative ring with identity. As in Subsection 1.4.2, we simply say "ring" instead of "commutative ring with identity". Recall that a ring is in particular an additive group, that is, an abelian group with respect to the binary operation of addition.

Definition 3.3.7 An *ideal* of a ring R is a subgroup J of the additive group R such that $ab \in J$ whenever $a \in R$ and $b \in J$.

Example 3.3.8 Every ring R has two trivial ideals, namely $J = \{0\}$ (called the *zero ideal*) and $J = R$. Now let $\mathbb{Z}$ be the ring of integers. We noted in Example 1.3.20 that for every $m \in \mathbb{N}$, the set $(m) := \{km : k \in \mathbb{Z}\}$ is a subgroup of the additive group $\mathbb{Z}$. It is obvious that (m) is in fact an ideal of $\mathbb{Z}$. Similarly, for every field F and every $f(x) \in F[x]$, the set $(f(x)) := \{g(x)f(x) : g(x) \in F[x]\}$ is an ideal of the polynomial ring $F[x]$. In general, for every ring R and every $b \in R$, the set $(b) := \{ab : a \in R\}$ is an ideal of R. An ideal of this type is called a *principal ideal*, and if we want to lay stress on the special role of the element b, then we say that (b) is the principal ideal generated by b.

Remark 3.3.9 Every ideal of $\mathbb{Z}$ is a principal ideal. If $J = \{0\}$, then J is the principal ideal generated by 0. If there are nonzero integers in J, then J contains positive integers by the property of being an additive group, and so there is a least positive integer m in J. Now it is easily seen that $J = (m)$. Obviously $(m) \subseteq J$, and if on the other hand $h \in J$ is arbitrary, then division with remainder yields $h = km + r$ with $k, r \in \mathbb{Z}$ and $0 \le r < m$; then $r = h - km \in J$ by the definition of an ideal, and so $r = 0$ by the minimality property of m, which shows that $h \in (m)$. Similarly, since there is a division with remainder in the polynomial ring $F[x]$ for an arbitrary field F, every ideal of $F[x]$ is principal.

In the theory of cyclic codes over $\mathbb{F}_q$, the notion of an ideal is applied to a residue class ring $\mathbb{F}_q[x]/(x^n - 1)$. We start out nice and easy with an example.

Example 3.3.10 Consider $\pi(C)$ in Example 3.3.5, viewed as a subset of the ring $R = \mathbb{F}_2[x]_{<4} = \mathbb{F}_2[x]/(x^4 - 1)$. We claim that $\pi(C)$ is an ideal of R. First we note that $\pi(C)$ is closed under addition since it is a vector space over $\mathbb{F}_2$, and so $\pi(C)$ is a subgroup of the additive group R. The verification of the remaining property in Definition 3.3.7 needs a bit of work. In a first step, we show that if $c(x) \in \pi(C)$, then also $xc(x) \in \pi(C)$. Since $\pi(C)$ has only four elements, this can be done by direct computation. Note that $x \cdot 0 = 0 \in \pi(C)$ and $x(1 + x^2) = x + x^3 \in \pi(C)$. Furthermore, $x(x + x^3) = x^2 + x^4 = x^2 + 1 \in \pi(C)$ since $x^4 = 1$ in R, and similarly

$$x(1 + x + x^2 + x^3) = x + x^2 + x^3 + x^4 = x + x^2 + x^3 + 1 \in \pi(C).$$

Thus, the first step is achieved. Now if $c(x) \in \pi(C)$, then $x^2c(x) = x(xc(x)) \in \pi(C)$ by what we have just proved, and similarly $x^3c(x) = x(x^2c(x)) \in \pi(C)$. Finally, an arbitrary $f(x) \in R$ is a sum of some of the monomials 1, x, x^2, x^3, hence $f(x)c(x)$ is a sum of some of the elements $c(x)$, $xc(x)$, $x^2c(x)$, $x^3c(x)$ of $\pi(C)$, and now the fact that $\pi(C)$ is closed under addition shows that $f(x)c(x) \in \pi(C)$. This completes the proof of the claim that $\pi(C)$ is an ideal of R.

Example 3.3.10 is an instance of a general fact, namely that for a cyclic code C over $\mathbb{F}_q$ of length n, the corresponding set $\pi(C)$ is an ideal of the residue class ring

$\mathbb{F}_q[x]/(x^n - 1)$. We recall that linear codes have, by definition, a dimension at least 1, and so the zero ideal mentioned in Example 3.3.8 cannot be of the form $\pi(C)$ for some cyclic code C.

Theorem 3.3.11 *Let $\pi : \mathbb{F}_q^n \to \mathbb{F}_q[x]/(x^n - 1)$ be the map defined in* (3.9)*. Then a subset C of $\mathbb{F}_q^n$ is a cyclic code if and only if $\pi(C)$ is a nonzero ideal of the residue class ring $\mathbb{F}_q[x]/(x^n - 1)$.*

Proof. We generalize the argument in Example 3.3.10. Let C be a cyclic code. Since π is a linear transformation, $\pi(C)$ is a subspace of $R = \mathbb{F}_q[x]/(x^n - 1)$ of dimension at least 1. In particular, $\pi(C)$ is closed under addition. Now let $c(x) = \sum_{j=0}^{n-1} c_j x^j \in \pi(C)$ be arbitrary. Then $\pi(\mathbf{c}) = c(x)$ with $\mathbf{c} = (c_0, c_1, \ldots, c_{n-1}) \in C$ and

$$\mathbf{c}^t = (c_{n-t}, \ldots, c_{n-1}, c_0, \ldots, c_{n-t-1}) \in C \qquad \text{for } 0 \le t \le n-1$$

since C is cyclic. Noting that $x^n = 1$ in R, we get

$$\begin{aligned}\pi(\mathbf{c}^t) &= c_{n-t} + \cdots + c_{n-1}x^{t-1} + c_0 x^t + \cdots + c_{n-t-1}x^{n-1} \\ &= c_{n-t}x^n + \cdots + c_{n-1}x^{n+t-1} + c_0 x^t + \cdots + c_{n-t-1}x^{n-1} \\ &= x^t(c_0 + \cdots + c_{n-t-1}x^{n-t-1} + c_{n-t}x^{n-t} + \cdots + c_{n-1}x^{n-1}) \\ &= x^t c(x).\end{aligned}$$

Therefore $x^t c(x) = \pi(\mathbf{c}^t) \in \pi(C)$ for $0 \le t \le n-1$. Now let $f(x) = \sum_{t=0}^{n-1} f_t x^t \in R$ with $f_0, f_1, \ldots, f_{n-1} \in \mathbb{F}_q$ be arbitrary. Then, recalling that $\pi(C)$ is a vector space over $\mathbb{F}_q$, we obtain $f(x)c(x) = \sum_{t=0}^{n-1} f_t x^t c(x) \in \pi(C)$, and so $\pi(C)$ is a nonzero ideal of R.

Conversely, suppose that $\pi(C)$ is a nonzero ideal of R. Since π is a bijective linear transformation, C is a subspace of $\mathbb{F}_q^n$ of dimension at least 1. By Remark 3.3.3, it remains to show that $\mathbf{c} \in C$ implies $\mathbf{c}^1 \in C$, and we can assume that $n \ge 2$. By the computation above, $\pi(\mathbf{c}^1) = xc(x) \in \pi(C)$ since $\pi(C)$ is an ideal of R, and so $\mathbf{c}^1 \in \pi^{-1}(\pi(C)) = C$. □

3.3.2 The generator polynomial

It is evident from Theorem 3.3.11 that in order to delve deeper into the structure of cyclic codes, we should study the nonzero ideals of $\mathbb{F}_q[x]/(x^n - 1)$.

Theorem 3.3.12 *Every ideal J of $\mathbb{F}_q[x]/(x^n - 1)$ is principal, and for every nonzero ideal J there exists a unique monic polynomial $g(x) \in \mathbb{F}_q[x]/(x^n - 1)$ such that J consists exactly of all multiples of $g(x)$. The polynomial $g(x)$ is a proper divisor of $x^n - 1$ in $\mathbb{F}_q[x]$.*

Proof. It suffices to consider a nonzero ideal J. Then there is a monic polynomial $g(x)$ of least degree in J. By Definition 3.3.7, every multiple of $g(x)$ is in J. We claim that conversely, if $f(x) \in J$, then $f(x)$ must be a multiple of $g(x)$. By the division algorithm, we can write $f(x) = a(x)g(x) + r(x)$ with $a(x), r(x) \in \mathbb{F}_q[x]$ and $\deg(r(x)) < \deg(g(x))$. Now $r(x) = f(x) - a(x)g(x) \in J$, and the minimality property of $g(x)$ implies that $r(x)$ is the zero polynomial. Thus, $f(x)$ is a multiple of $g(x)$, and so J consists exactly of all multiples of $g(x)$.

If $g_1(x) \in \mathbb{F}_q[x]/(x^n - 1)$ is an arbitrary monic polynomial such that J consists exactly of all multiples of $g_1(x)$, then from $g(x), g_1(x) \in J$ we infer that $g(x)$ divides $g_1(x)$ and $g_1(x)$ divides $g(x)$. Since $g(x)$ and $g_1(x)$ are both monic, this implies $g_1(x) = g(x)$ and shows the uniqueness of $g(x)$.

Note that $x^n - 1$ is the zero element of $\mathbb{F}_q[x]/(x^n - 1)$, thus it belongs to J and is therefore a multiple of $g(x)$. Moreover, $\deg(g(x)) < n$ by construction, and so $g(x)$ is a proper divisor of $x^n - 1$ in $\mathbb{F}_q[x]$. All statements in the theorem have now been proved. □

Definition 3.3.13 For a nonzero ideal J of $\mathbb{F}_q[x]/(x^n - 1)$, the uniquely determined polynomial $g(x)$ in Theorem 3.3.12 is called the *generator polynomial* of J. For a cyclic code C, the generator polynomial of the nonzero ideal $\pi(C)$ is called the *generator polynomial* of C.

Example 3.3.14 We determine the generator polynomials $g(x)$ of the cyclic codes C listed in Example 3.3.4. By the proof of Theorem 3.3.12, in each case it suffices to find the monic polynomial $g(x)$ of least degree in $\pi(C)$.
(i) If C is as in Example 3.3.4(i), then it was shown in Example 3.3.5 that

$$\pi(C) = \{0,\ 1 + x^2,\ x + x^3,\ 1 + x + x^2 + x^3\}.$$

Therefore $g(x) = x^2 + 1$. Note that $g(x)$ divides $x^4 - 1 = x^4 + 1$ in $\mathbb{F}_2[x]$ since $x^4 + 1 = (x^2 + 1)^2$.
(ii) For the repetition code C over $\mathbb{F}_q$ of length n, it is clear that

$$\pi(C) = \{a(1 + x + x^2 + \cdots + x^{n-1}) : a \in \mathbb{F}_q\}.$$

Therefore $g(x) = x^{n-1} + \cdots + x^2 + x + 1$. Again, $g(x)$ divides $x^n - 1$ in $\mathbb{F}_q[x]$ since

$$x^n - 1 = (x^{n-1} + \cdots + x^2 + x + 1)(x - 1).$$

(iii) For $C = \mathbb{F}_q^n$ we get $\pi(C) = \mathbb{F}_q[x]/(x^n - 1)$, and so $g(x) = 1$.

Theorem 3.3.15 *There is a one-to-one correspondence between the nonzero ideals of $\mathbb{F}_q[x]/(x^n - 1)$, and so of the cyclic codes over $\mathbb{F}_q$ of length n, and the monic proper divisors of $x^n - 1$ in $\mathbb{F}_q[x]$.*

Proof. To each nonzero ideal of $\mathbb{F}_q[x]/(x^n - 1)$ there corresponds a unique monic proper divisor of $x^n - 1$ in $\mathbb{F}_q[x]$, according to Theorem 3.3.12. On the other hand, if $g(x)$ is a monic proper divisor of $x^n - 1$ in $\mathbb{F}_q[x]$, then the multiples of $g(x)$ form a nonzero ideal of $\mathbb{F}_q[x]/(x^n - 1)$. Furthermore, Theorem 3.3.12 implies that different nonzero ideals of $\mathbb{F}_q[x]/(x^n - 1)$ correspond to different monic proper divisors of $x^n - 1$ in $\mathbb{F}_q[x]$. □

Example 3.3.16 We determine all cyclic codes over $\mathbb{F}_3$ of length 4. In view of Theorem 3.3.15, this is equivalent to finding all monic proper divisors of $x^4 - 1 = x^4 + 2$ in $\mathbb{F}_3[x]$. We start from the canonical factorization

$$x^4 + 2 = (x + 1)(x + 2)(x^2 + 1)$$

into monic irreducible polynomials over $\mathbb{F}_3$. Therefore the monic proper divisors of $x^4 + 2$ in $\mathbb{F}_3[x]$ are given by

$$1,\ x + 1,\ x + 2,\ (x + 1)(x + 2),\ x^2 + 1,\ (x + 1)(x^2 + 1),\ (x + 2)(x^2 + 1).$$

Thus, there are exactly seven different cyclic codes over $\mathbb{F}_3$ of length 4, each having a generator polynomial from the list of seven polynomials above. Let us explicitly describe, for instance, the cyclic code C with generator polynomial $g(x) = x^2 + 1$. By computing all multiples of $g(x)$ in $\mathbb{F}_3[x]/(x^4 + 2)$, we get the ideal

$$\begin{aligned}\pi(C) = \{&0,\ 1 + x^2,\ 2 + 2x^2,\ x + x^3,\ 1 + x + x^2 + x^3,\ 2 + x + 2x^2 + x^3,\\ &2x + 2x^3,\ 2 + 2x + 2x^2 + 2x^3,\ 1 + 2x + x^2 + 2x^3\}.\end{aligned}$$

By applying the inverse π^{-1} of the map π to each element of $\pi(C)$, we obtain the cyclic code

$$C = \{(0000),\ (1010),\ (2020),\ (0101),\ (1111),\ (2121),\ (0202),\ (2222),\ (1212)\}$$

over $\mathbb{F}_3$ of length 4.

An important issue for a cyclic code, as for any linear code, is the determination of the dimension of the code. Since a cyclic code is often given via its generator polynomial, the question is how we can read off the dimension from the generator polynomial. The following theorem provides the answer.

Theorem 3.3.17 *Let $g(x)$ be the generator polynomial of the cyclic code C over $\mathbb{F}_q$ of length n. Then*

$$\dim(C) = n - \deg(g(x)).$$

Proof. In view of Proposition 3.2.6 and since π is a bijection, it suffices to show that the ideal $J = \pi(C)$ corresponding to C has exactly $q^{n-\deg(g(x))}$ elements. Recall that J consists of the multiples of $g(x)$ in $\mathbb{F}_q[x]/(x^n - 1)$. Let $f(x)g(x)$ be such a multiple. With $m = \deg(g(x))$ and by the division algorithm, we can write

$$f(x) = a(x)\frac{x^n - 1}{g(x)} + r(x)$$

with $a(x), r(x) \in \mathbb{F}_q[x]$ and $\deg(r(x)) < n - m$. Then

$$f(x)g(x) = a(x)(x^n - 1) + r(x)g(x) = r(x)g(x)$$

in $\mathbb{F}_q[x]/(x^n - 1)$ since $x^n - 1 = 0$ in $\mathbb{F}_q[x]/(x^n - 1)$. Thus,

$$J = \{r(x)g(x) : \deg(r(x)) < n - m\}.$$

Now we claim that distinct choices of $r(x)$ yield distinct elements of J. So take $r_1(x)$ and $r_2(x)$ with $\deg(r_1(x)) < n - m$ and $\deg(r_2(x)) < n - m$ such that $r_1(x)g(x) = r_2(x)g(x)$ in $\mathbb{F}_q[x]/(x^n - 1)$. Then $x^n - 1$ divides $r_1(x)g(x) - r_2(x)g(x) = (r_1(x) - r_2(x))g(x)$, and so $(x^n - 1)/g(x)$ divides $r_1(x) - r_2(x)$. By comparing degrees, we see that $r_1(x) = r_2(x)$, hence the claim is demonstrated. It follows that the number of elements of J is equal to the number of choices for $r(x)$, which is $q^{n-m} = q^{n-\deg(g(x))}$. □

Example 3.3.18 Consider the cyclic code C over $\mathbb{F}_3$ in Example 3.3.16. This code has length $n = 4$ and generator polynomial $g(x) = x^2 + 1$. Hence $\dim(C) = n - \deg(g(x)) = 2$ by Theorem 3.3.17. A basis of C is formed by the codewords (1 0 1 0) and (0 1 0 1).

3.3.3 Generator matrix

We recall from Subsection 3.2.4 that every linear code has a generator matrix. If C is a cyclic code over $\mathbb{F}_q$ of length n and with $\dim(C) = k$, then a generator matrix of C must be a $k \times n$ matrix over $\mathbb{F}_q$ whose row vectors form a basis of C. If C is given via its generator polynomial $g(x) \in \mathbb{F}_q[x]$, then $\deg(g(x)) = n - k$ by Theorem 3.3.17. Let us write

$$g(x) = g_0 + g_1 x + \cdots + g_{n-k}x^{n-k} \in \mathbb{F}_q[x] \qquad (3.10)$$

with $g_0, g_1, \ldots, g_{n-k} \in \mathbb{F}_q$ and $g_{n-k} = 1$. Then a generator matrix of C can be immediately derived from $g(x)$.

Theorem 3.3.19 *Let $g(x)$ in* (3.10) *be the generator polynomial of a cyclic code $C \subseteq \mathbb{F}_q^n$ with* $\deg(g(x)) = n - k$. *Then the $k \times n$ matrix*

$$G = \begin{pmatrix} g_0 & g_1 & \cdot\ \cdot\ \cdot & g_{n-k} & 0 & 0\ 0\ .\ .\ . & 0 \\ 0 & g_0 & g_1\ \cdot\ \cdot & \cdot & g_{n-k} & 0\ 0\ .\ .\ . & 0 \\ \cdot & & & & & & \cdot \\ \cdot & & & & & & \cdot \\ \cdot & & & & & & \cdot \\ 0 & 0 & .\ .\ . & g_0 & g_1 & .\ .\ .\ .\ . & g_{n-k} \end{pmatrix}$$

over $\mathbb{F}_q$ is a generator matrix of C.

Proof. Note that

$$\pi^{-1}(g(x)) = (g_0\ g_1\ .\ .\ .\ g_{n-k}\ 0\ 0\ 0\ .\ .\ .\ 0)$$

belongs to C. Since C is cyclic, all cyclic shifts of this vector are codewords in C. In particular, all row vectors of G belong to C. Since $g_{n-k} = 1$, it is clear that the row vectors of G are linearly independent over $\mathbb{F}_q$. The number of row vectors of G is $k = \dim(C)$, and so the row vectors of G form a basis of C. □

Example 3.3.20 Let C be the ternary cyclic code in Example 3.3.16 with generator polynomial $g(x) = 1 + x^2 \in \mathbb{F}_3[x]$. Then a generator matrix of C is given by

$$G = \begin{pmatrix} 1\ 0\ 1\ 0 \\ 0\ 1\ 0\ 1 \end{pmatrix}.$$

We observe that $g_0 \neq 0$ in (3.10) since $g(x)$ divides $x^n - 1$ by Theorem 3.3.12. Therefore the matrix G in Theorem 3.3.19 can be transformed into standard form (see Definition 3.2.25) by elementary row operations. Consequently, every cyclic code has a generator matrix in standard form, whereas in general a linear code need not have a generator matrix in standard form (see Example 3.2.26). We have seen in Subsection 3.2.4 that linear codes with a generator matrix in standard form have fast encoders and decoders.

It is a natural question to ask whether the unique generator matrix in standard form of a cyclic code can be derived from its generator polynomial by polynomial manipulations. This is indeed the case, and one proceeds as follows. Let $C \subseteq \mathbb{F}_q^n$ be a cyclic code with $\dim(C) = k$. The case $k = n$ is trivial, and so we can assume that $k < n$. Let $g(x) \in \mathbb{F}_q[x]$ with $\deg(g(x)) = n - k$ be the generator polynomial of C. For every integer $j \geq 0$, we can use the division algorithm to write

$$x^j = a_j(x)g(x) + r_j(x) \tag{3.11}$$

with $a_j(x), r_j(x) \in \mathbb{F}_q[x]$ and $\deg(r_j(x)) < n - k$. Let the map $\sigma : \mathbb{F}_q^{n-k} \to \mathbb{F}_q[x]_{<n-k}$ be as in (3.9), but with n replaced by $n - k$.

Theorem 3.3.21 *Let $C \subseteq \mathbb{F}_q^n$ be a cyclic code with* $\dim(C) = k < n$. *Let*

$$G = (I_k \mid -T),$$

where the $k \times (n-k)$ matrix T over $\mathbb{F}_q$ is defined as follows: for $1 \leq i \leq k$, the ith row is $\sigma^{-1}(r_{n-k-1+i}(x))$ with the notation in (3.11). *Then G is the unique generator matrix of C in standard form.*

Proof. It suffices to show that the rows of G are codewords in C. It follows from (3.11) that $x^j - r_j(x)$ is a multiple of $g(x)$, and so

$$c_j(x) := x^k(x^j - r_j(x)) \in \pi(C) \subseteq \mathbb{F}_q[x]/(x^n - 1)$$

for all integers $j \geq 0$. Using the fact that $x^n = 1$ in $\mathbb{F}_q[x]/(x^n-1)$, we deduce that

$$\begin{aligned} c_{n-k-1+i}(x) &= x^k(x^{n-k-1+i} - r_{n-k-1+i}(x)) = x^{n-1+i} - x^k r_{n-k-1+i}(x) \\ &= x^{i-1} - x^k r_{n-k-1+i}(x) \in \pi(C) \end{aligned}$$

for $1 \leq i \leq k$. This implies that $\pi^{-1}(c_{n-k-1+i}(x))$, the ith row of G, is a codeword in C. □

Example 3.3.22 Since $g(x) = x^3 + x^2 + 1 \in \mathbb{F}_2[x]$ divides $x^7 - 1 = x^7 + 1 \in \mathbb{F}_2[x]$, there exists a binary cyclic code C of length 7 with generator polynomial $g(x)$. Note that $\dim(C) = 4$ by Theorem 3.3.17. In order to find the generator matrix G of C in standard form, it suffices, by Theorem 3.3.21, to determine the polynomials $r_3(x), r_4(x), r_5(x), r_6(x)$. These are obtained from (3.11) by computing x^3, x^4, x^5, x^6 modulo $g(x)$. This computation in the residue class ring $\mathbb{F}_q[x]/(g(x))$ can be carried out using congruences modulo $g(x)$ (see Subsection 1.4.3), and this yields

$$\begin{aligned} x^3 &\equiv x^2 + 1 \pmod{g(x)}, \\ x^4 &\equiv x^3 + x \equiv x^2 + x + 1 \pmod{g(x)}, \\ x^5 &\equiv x^3 + x^2 + x \equiv x + 1 \pmod{g(x)}, \\ x^6 &\equiv x^2 + x \pmod{g(x)}. \end{aligned}$$

Therefore $r_3(x) = 1 + x^2$, $r_4(x) = 1 + x + x^2$, $r_5(x) = 1 + x$, $r_6(x) = x + x^2$. From the coefficients of these polynomials we obtain the matrix T in Theorem 3.3.21, and the final result is the generator matrix

$$G = \begin{pmatrix} 1\,0\,0\,0\,1\,0\,1 \\ 0\,1\,0\,0\,1\,1\,1 \\ 0\,0\,1\,0\,1\,1\,0 \\ 0\,0\,0\,1\,0\,1\,1 \end{pmatrix}.$$

3.3.4 Dual code and parity-check matrix

In Definition 3.2.33, we introduced for every nontrivial linear code C, that is, for every linear $[n, k]$ code C with $1 \leq k \leq n-1$, its dual code $C^{\perp}$. In the following, we study the dual code of a nontrivial cyclic code.

Proposition 3.3.23 *The dual code of a nontrivial cyclic code is again cyclic.*

Proof. A nontrivial cyclic code C automatically has length $n \geq 2$, and so by Remark 3.3.3 it suffices to show that $\mathbf{u} = (u_0, u_1, \ldots, u_{n-1}) \in C^{\perp}$ implies $\mathbf{u}^1 \in C^{\perp}$. If $\mathbf{c} = (c_0, c_1, \ldots, c_{n-1}) \in C$, then

$$\mathbf{u}^1 \cdot \mathbf{c} = u_{n-1}c_0 + u_0c_1 + \cdots + u_{n-2}c_{n-1} = \mathbf{u} \cdot \mathbf{c}^{n-1} = 0$$

since $\mathbf{u} \in C^{\perp}$ and $\mathbf{c}^{n-1} \in C$. Hence $\mathbf{u}^1 \in C^{\perp}$ as required. □

Since the dual code $C^{\perp}$ of a nontrivial cyclic code C is cyclic, $C^{\perp}$ has a uniquely determined generator polynomial. How is the generator polynomial of $C^{\perp}$ related to the generator polynomial of C? The answer is given in the following theorem. First we need a simple definition.

Definition 3.3.24 Let $h(x) \in \mathbb{F}_q[x]$ be a polynomial of degree $k \geq 1$. Then the *reciprocal polynomial* $h^*(x)$ of $h(x)$ is defined by $h^*(x) = x^k h(1/x) \in \mathbb{F}_q[x]$.

Example 3.3.25 For $h(x) = x^3 + 2x^2 + 4x + 3 \in \mathbb{F}_5[x]$, its reciprocal polynomial is

$$\begin{aligned} h^*(x) &= x^3 h(1/x) = x^3[(1/x)^3 + 2(1/x)^2 + 4(1/x) + 3] \\ &= 3x^3 + 4x^2 + 2x + 1. \end{aligned}$$

In general, the reciprocal polynomial of $h(x) \in \mathbb{F}_q[x]$ is obtained by reading the coefficients of $h(x)$ in reverse order.

Theorem 3.3.26 *Let C be a nontrivial cyclic $[n, k]$ code over $\mathbb{F}_q$ with generator polynomial $g(x)$. Put $h(x) = (x^n - 1)/g(x) \in \mathbb{F}_q[x]$. Then the generator polynomial of the dual code $C^{\perp}$ is $h_0^{-1}h^*(x)$, where h_0 is the constant term of $h(x)$.*

Proof. First we note that $g(x)$ divides $x^n - 1$ by Theorem 3.3.12, and so $h(x) = (x^n - 1)/g(x)$ is indeed a polynomial over $\mathbb{F}_q$. Let $m(x) \in \mathbb{F}_q[x]$ be the generator polynomial of $C^{\perp}$. Then by Theorems 3.3.17 and 3.2.34,

$$\deg(m(x)) = n - \dim(C^{\perp}) = n - (n-k) = k.$$

Note that $\deg(h(x)) = n - \deg(g(x)) = k$ by Theorem 3.3.17. Furthermore, we observe that $h_0 \neq 0$ since $h(x)$ divides $x^n - 1$, and so $\deg(h^*(x)) = k$. Thus, $h_0^{-1}h^*(x)$ is a monic polynomial of degree k. If we can show that $h^*(x) \in \pi(C^{\perp})$, then we can conclude that $m(x) = h_0^{-1}h^*(x)$ since there is exactly one monic polynomial of degree k in the ideal $\pi(C^\perp)$.

Let $g(x)$ be as in (3.10) and let

$$\mathbf{g} = (g_0, g_1, \dots, g_{n-1}) \in \mathbb{F}_q^n$$

be the first row vector of the generator matrix G of C in Theorem 3.3.19, where $g_j = 0$ for $n-k < j \le n-1$. Let $h(x) = \sum_{j=0}^{n-1} h_j x^j$ and

$$\mathbf{u} = (h_{n-1}, h_{n-2}, \dots, h_0) \in \mathbb{F}_q^n,$$

where $h_j = 0$ for $k < j \le n-1$. Note that $g(x)h(x) = x^n - 1$ in $\mathbb{F}_q[x]$ by the definition of $h(x)$. By comparing the coefficients of x^{n-1} in this identity, we get

$$g_0 h_{n-1} + g_1 h_{n-2} + \cdots + g_{n-1} h_0 = 0,$$

and so $\mathbf{g} \cdot \mathbf{u} = 0$. Similarly, by comparing the coefficients of x^{n-t} for $1 \le t \le k$, we get $\mathbf{g}^{t-1} \cdot \mathbf{u} = 0$ for $1 \le t \le k$ with the notation in Definition 3.3.1, and so $\mathbf{u} \in C^\perp$. Since $C^\perp$ is cyclic by Proposition 3.3.23, we have $\mathbf{u}^{k+1} \in C^\perp$. Now $\pi(\mathbf{u}^{k+1}) = h^*(x)$, and so $h^*(x) \in \pi(C^\perp)$ as desired. □

Remark 3.3.27 For a nontrivial cyclic code C, it is convenient to use the terminology *parity-check polynomial* of C for the generator polynomial $h_0^{-1}h^*(x)$ of $C^\perp$ in Theorem 3.3.26. The parity-check polynomial of C divides again $x^n - 1$, where n is the length of C.

Example 3.3.28 Let C be the ternary cyclic code of length 8 with generator polynomial $g(x) = x^2 + 1 \in \mathbb{F}_3[x]$. Then

$$h(x) = (x^8 - 1)/g(x) = (x^8 - 1)/(x^2 + 1) = x^6 + 2x^4 + x^2 + 2 \in \mathbb{F}_3[x].$$

Furthermore, $h^*(x) = 2x^6 + x^4 + 2x^2 + 1$ and $h_0 = 2$, and so by Theorem 3.3.26 the generator polynomial of the dual code $C^\perp$, or in other words the parity-check polynomial of C, is $h_0^{-1}h^*(x) = 2h^*(x) = x^6 + 2x^4 + x^2 + 2 \in \mathbb{F}_3[x]$.

We are now in a position to determine a parity-check matrix of a given nontrivial cyclic code C over $\mathbb{F}_q$ of length n. We note that the generator polynomial of the dual code $C^\perp$ is obtained from Theorem 3.3.26, and so a generator matrix of $C^\perp$ can be set up by Theorem 3.3.19. But now this generator matrix of $C^\perp$ serves by Definition 3.2.36 as a parity-check matrix of C.

Example 3.3.29 Let C be the ternary cyclic code in Example 3.3.28. Then the generator polynomial of $C^\perp$ is $x^6+2x^4+x^2+2 \in \mathbb{F}_3[x]$. The corresponding generator matrix of $C^\perp$ is

$$H = \begin{pmatrix} 2 & 0 & 1 & 0 & 2 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 & 0 & 2 & 0 & 1 \end{pmatrix}$$

according to Theorem 3.3.19, and so H is a parity-check matrix of C.

3.3.5 Cyclic codes from roots

Cyclic codes over $\mathbb{F}_q$ can be introduced also by means of roots of polynomials over $\mathbb{F}_q$. Let $\alpha_1, \ldots, \alpha_s$ be nonzero elements in some finite extension field of $\mathbb{F}_q$. For $i = 1, \ldots, s$, let $m_i(x) \in \mathbb{F}_q[x]$ be the minimal polynomial of α_i over $\mathbb{F}_q$. Put

$$g(x) = \operatorname{lcm}(m_1(x), \ldots, m_s(x)) \in \mathbb{F}_q[x]. \tag{3.12}$$

Let n be a positive integer such that $\alpha_i^n = 1$ for $1 \leq i \leq s$. Then $g(x)$ divides $x^n - 1$ in $\mathbb{F}_q[x]$ by Proposition 1.4.38. If we assume that $\deg(g(x)) < n$, then $g(x)$ is the generator polynomial of a cyclic code over $\mathbb{F}_q$ of length n. The codewords in this cyclic code can be characterized as follows.

Theorem 3.3.30 *Let $C \subseteq \mathbb{F}_q^n$ be the cyclic code with the generator polynomial $g(x)$ in* (3.12) *satisfying* $\deg(g(x)) < n$ *and let $\mathbf{v} \in \mathbb{F}_q^n$. Then $\mathbf{v} \in C$ if and only if the polynomial $v = \pi(\mathbf{v}) \in \mathbb{F}_q[x]$ given by* (3.9) *satisfies $v(\alpha_i) = 0$ for $1 \leq i \leq s$.*

Proof. If $\mathbf{v} \in C$, then $g(x)$ divides $v(x)$ in $\mathbb{F}_q[x]$ by the definition of the generator polynomial of C. For each $i = 1, \ldots, s$, the polynomial $m_i(x)$ divides $g(x)$ in $\mathbb{F}_q[x]$, and so $m_i(x)$ divides $v(x)$ in $\mathbb{F}_q[x]$. Now $m_i(\alpha_i) = 0$, and thus $v(\alpha_i) = 0$. Conversely, if $v(\alpha_i) = 0$ for $1 \leq i \leq s$, then $m_i(x)$ divides $v(x)$ in $\mathbb{F}_q[x]$ for $1 \leq i \leq s$ by Proposition 1.4.38, and so $g(x)$ divides $v(x)$ in $\mathbb{F}_q[x]$. This shows that $\mathbf{v} \in C$. □

Example 3.3.31 Let $\alpha_1 \in \mathbb{F}_4$ be a root of the irreducible polynomial $x^2 + x + 1$ over $\mathbb{F}_2$ and let $\alpha_2 \in \mathbb{F}_8$ be a root of the irreducible polynomial $x^3 + x + 1$ over $\mathbb{F}_2$. Then $\alpha_1^3 = 1$ and $\alpha_2^7 = 1$, hence $\alpha_1^{21} = \alpha_2^{21} = 1$. Thus, from α_1 and α_2 we obtain a cyclic code C over $\mathbb{F}_2$ of length 21. The generator polynomial of C is

$$g(x) = \operatorname{lcm}(x^2 + x + 1, x^3 + x + 1) = (x^2 + x + 1)(x^3 + x + 1) = x^5 + x^4 + 1.$$

The cyclic code C has dimension 16 by Theorem 3.3.17. A polynomial $v(x) \in \mathbb{F}_2[x]$ belongs to the ideal $\pi(C)$ if and only if $v(\alpha_1) = v(\alpha_2) = 0$.

There is no easy general formula for the minimum distance of a cyclic code, but there are various results that yield lower bounds on the minimum distance. The following considerations lead to useful tools for establishing such bounds.

Let $\mathbb{F}_q$ be a finite field and let $n \geq 2$ be an integer with $\gcd(n, q) = 1$. Then there exists a finite extension field of $\mathbb{F}_q$ containing a primitive nth root of unity. Indeed, the condition $\gcd(n, q) = 1$ implies by Theorem 1.2.15 that there exists a positive integer k with $q^k \equiv 1 \pmod{n}$. Let β be a primitive element of $\mathbb{F}_{q^k}$ and put $\gamma = \beta^{(q^k-1)/n}$. Then γ is an element of order n in the multiplicative group $\mathbb{F}_{q^k}^*$; in other words, γ is a primitive nth root of unity.

We recall that to each $\mathbf{v} \in \mathbb{F}_q^n$ we can associate a polynomial $v = \pi(\mathbf{v}) \in \mathbb{F}_q[x]$ according to (3.9). Here is another polynomial associated to $\mathbf{v}$.

Definition 3.3.32 Let $n \geq 2$ be an integer with $\gcd(n, q) = 1$ and let γ be a primitive nth root of unity in a finite extension field of $\mathbb{F}_q$. Then for every $\mathbf{v} \in \mathbb{F}_q^n$, the *Mattson-Solomon polynomial* $M_{\mathbf{v}}(x)$ of $\mathbf{v}$ is defined by

$$M_{\mathbf{v}}(x) = \sum_{j=1}^{n} v(\gamma^j) x^{n-j},$$

where $v(x) \in \mathbb{F}_q[x]$ is the polynomial corresponding to $\mathbf{v}$ according to (3.9).

Note that if $\gamma \in \mathbb{F}_{q^k}$, then $M_{\mathbf{v}}(x)$ is a polynomial over $\mathbb{F}_{q^k}$. The Mattson-Solomon polynomial may depend also on the specific choice of γ, but we think of γ as being fixed and thus suppress this dependence in the notation. The coordinates of $\mathbf{v}$ can be recovered from $M_{\mathbf{v}}(x)$ in the following way.

Lemma 3.3.33 *Let $n \geq 2$ be an integer with $\gcd(n, q) = 1$ and let γ be a primitive nth root of unity in a finite extension field of $\mathbb{F}_q$. If $\mathbf{v} = (v_0, v_1, \ldots, v_{n-1}) \in \mathbb{F}_q^n$, then*

$$v_i = n^{-1} M_{\mathbf{v}}(\gamma^i) \qquad \text{for } i = 0, 1, \ldots, n-1,$$

where n^{-1} is the multiplicative inverse of n considered as an element of the prime subfield of $\mathbb{F}_q$.

Proof. For $i = 0, 1, \ldots, n-1$, we obtain

$$\begin{aligned} M_{\mathbf{v}}(\gamma^i) &= \sum_{j=1}^{n} v(\gamma^j)\gamma^{i(n-j)} = \sum_{j=1}^{n} v(\gamma^j)\gamma^{-ij} \\ &= \sum_{j=1}^{n} \gamma^{-ij} \sum_{h=0}^{n-1} v_h \gamma^{hj} = \sum_{h=0}^{n-1} v_h \sum_{j=0}^{n-1} \gamma^{(h-i)j} = n v_i, \end{aligned}$$

since the formula for geometric sums shows that $\sum_{j=0}^{n-1} \gamma^{(h-i)j} = 0$ for $h \in \{0, 1, \ldots, n-1\}$ with $h \neq i$. □

We are now ready to prove a classical lower bound on the minimum distance of cyclic codes.

Theorem 3.3.34 *Let $C \subseteq \mathbb{F}_q^n$ be a cyclic code with $n \geq 2$, $\gcd(n, q) = 1$, and generator polynomial $g(x)$. Let γ be a primitive nth root of unity in a finite extension field of $\mathbb{F}_q$. Assume that there exist integers b and d with $b \geq 0$ and $2 \leq d \leq n$ such that $g(\gamma^{b+i}) = 0$ for $0 \leq i \leq d-2$. Then the minimum distance of C is at least d.*

Proof. Let $\mathbf{v} \in \mathbb{F}_q^n$ be a nonzero codeword in C. Then the corresponding polynomial $v(x) \in \mathbb{F}_q[x]$ is nonzero and satisfies $\deg(v(x)) < n$. Since the n distinct elements γ^j, $j = 1, \ldots, n$, cannot all be roots of $v(x)$, the Mattson-Solomon polynomial $M_{\mathbf{v}}(x)$ is nonzero. By multiplying $M_{\mathbf{v}}(x)$, if necessary, by a suitable power of x modulo x^n-1, we can assume that $b=1$. Then $g(\gamma^j)=0$ for $1 \le j \le d-1$ by the hypothesis, and since $g(x)$ divides $v(x)$ in $\mathbb{F}_q[x]$, we deduce that $v(\gamma^j)=0$ for $1 \le j \le d-1$. It follows then from Definition 3.3.32 that $\deg(M_{\mathbf{v}}(x)) \le n-d$. By Lemma 3.3.33 the Hamming weight $w(\mathbf{v})$ satisfies $w(\mathbf{v})=n-r$, where r is the number of nth roots of unity that are roots of $M_{\mathbf{v}}(x)$. Now trivially $r \le \deg(M_{\mathbf{v}}(x))$, and so $r \le n-d$. This implies $w(\mathbf{v})=n-r \ge d$, and since this holds for every nonzero $\mathbf{v} \in C$, the desired result follows from Theorem 3.2.14. □

Remark 3.3.35 If you are familiar with determinants, then you will appreciate the following alternative proof of Theorem 3.3.34. We proceed by contradiction and suppose that there exists a nonzero codeword $\mathbf{c} \in C$ with Hamming weight $w=w(\mathbf{c})<d$. Let $u=\pi(\mathbf{c}) \in \mathbb{F}_q[x]$ be the corresponding polynomial. Then $g(x)$ divides $u(x)$ in $\mathbb{F}_q[x]$, and so $u(\gamma^{b+i})=0$ for $0 \le i \le d-2$. Since $w=w(\mathbf{c})$, we can write

$$u(x)=\sum_{j=1}^{w} u_j x^{a_j}$$

with $u_j \in \mathbb{F}_q^*$ for $1 \le j \le w$ and integers $0 \le a_1 < a_2 < \cdots < a_w < n$. The property $u(\gamma^{b+i})=0$ for $0 \le i \le d-2$ implies that $u(\gamma^{b+i})=0$ for $0 \le i \le w-1$. This can be put in the form $K\mathbf{u}^\top=\mathbf{0}^\top$, where $\mathbf{u}=(u_1,\dots,u_w)$ and K is the $w \times w$ matrix

$$K=\begin{pmatrix} \gamma^{a_1 b} & \gamma^{a_2 b} & \dots \gamma^{a_w b} \\ \gamma^{a_1(b+1)} & \gamma^{a_2(b+1)} & \dots \gamma^{a_w(b+1)} \\ \vdots & \vdots & \vdots \\ \gamma^{a_1(b+w-1)} & \gamma^{a_2(b+w-1)} & \dots \gamma^{a_w(b+w-1)} \end{pmatrix}.$$

A basic property of determinants yields $\det(K)=\gamma^{(a_1+a_2+\cdots+a_w)b}\det(L)$, where L is the Vandermonde matrix

$$L=\begin{pmatrix} 1 & 1 & \dots 1 \\ \gamma^{a_1} & \gamma^{a_2} & \dots \gamma^{a_w} \\ \vdots & \vdots & \vdots \\ \gamma^{a_1(w-1)} & \gamma^{a_2(w-1)} & \dots \gamma^{a_w(w-1)} \end{pmatrix}.$$

Now $\gamma^{a_1},\gamma^{a_2},\dots,\gamma^{a_w}$ are distinct since γ is a primitive nth root of unity, hence $\det(L) \ne 0$, and so $\det(K) \ne 0$. Thus, $K\mathbf{u}^\top=\mathbf{0}^\top$ implies that $\mathbf{u}=\mathbf{0}$. This contradiction completes the alternative proof of Theorem 3.3.34.

Example 3.3.36 Let C be the binary cyclic code in Example 3.3.31. With a suitable primitive 21st root of unity $\gamma \in \mathbb{F}_{64}$, we can take $\alpha_1=\gamma^7$ and $\alpha_2=\gamma^3$ in Example 3.3.31. Then by Proposition 1.4.47, the roots of the generator polynomial

$g(x) = x^5 + x^4 + 1 \in \mathbb{F}_2[x]$ are $\gamma^7, \gamma^{14}, \gamma^3, \gamma^6$, and γ^{12}. Thus, with $b = 6$ and $d = 3$ we get $g(\gamma^{b+i}) = 0$ for $0 \le i \le d-2$. It follows then from Theorem 3.3.34 that $d(C) \ge 3$. Since $g(x)$ corresponds to a codeword in C of Hamming weight 3, we conclude that $d(C) = 3$.

3.3.6 Irreducible cyclic codes

Now we consider a special family of cyclic codes that allow a nice explicit description of the codewords. As in the discussion prior to Definition 3.3.32, we take a finite field $\mathbb{F}_q$ and let $n \ge 2$ be an integer with $\gcd(n, q) = 1$. Let k be the least positive integer such that $q^k \equiv 1 \pmod{n}$. Then there exists a primitive nth root of unity $\gamma \in \mathbb{F}_{q^k}$. Let $f(x) \in \mathbb{F}_q[x]$ be the minimal polynomial of γ over $\mathbb{F}_q$. Then $f(x)$ is a monic irreducible polynomial over $\mathbb{F}_q$ of degree k. We note that $f(x)$ divides $x^n - 1$ in $\mathbb{F}_q[x]$ since $\gamma^n = 1$ (see Proposition 1.4.38). Hence there exists a cyclic $[n, k]$ code C over $\mathbb{F}_q$ with parity-check polynomial $f(x)$ (compare with Remark 3.3.27). Such a cyclic code is called an *irreducible cyclic code* with parity-check polynomial $f(x)$. Here is the promised explicit description of the codewords in C. We use the trace map for finite fields that we introduced in Definition 1.4.48 and the basic properties of the trace map in Theorem 1.4.50.

Theorem 3.3.37 *Let $n \ge 2$ be an integer with $\gcd(n, q) = 1$ and let C be the irreducible cyclic $[n, k]$ code over $\mathbb{F}_q$ with parity-check polynomial $f(x)$, where $f(x)$ is the minimal polynomial of the primitive nth root of unity $\gamma \in \mathbb{F}_{q^k}$ over $\mathbb{F}_q$. Then the codewords in C are exactly the words*

$$\mathbf{c}(\theta) = (\mathrm{Tr}(\theta), \mathrm{Tr}(\theta\gamma), \mathrm{Tr}(\theta\gamma^2), \ldots, \mathrm{Tr}(\theta\gamma^{n-1})) \in \mathbb{F}_q^n,$$

where θ runs through the finite field $\mathbb{F}_{q^k}$ and Tr denotes the trace map from $\mathbb{F}_{q^k}$ onto $\mathbb{F}_q$.

Proof. By definition, $f(x) = \sum_{j=0}^{k} f_j x^j \in \mathbb{F}_q[x]$ with all $f_j \in \mathbb{F}_q$ is the generator polynomial of the dual code $C^\perp$. A parity-check matrix of C is a generator matrix of $C^\perp$, and so a parity-check matrix H of C is obtained from Theorem 3.3.19 as the $(n-k) \times n$ matrix

$$H = \begin{pmatrix} f_0 & f_1 & \ldots & f_k & 0 & 0 & 0 & \ldots & 0 \\ 0 & f_0 & f_1 & \ldots & f_k & 0 & 0 & \ldots & 0 \\ \cdot & & & & & & & & \cdot \\ \cdot & & & & & & & & \cdot \\ \cdot & & & & & & & & \cdot \\ 0 & 0 & \ldots & f_0 & f_1 & \ldots & & & f_k \end{pmatrix}$$

over $\mathbb{F}_q$. For $\mathbf{v} = (v_0, v_1, \ldots, v_{n-1}) \in \mathbb{F}_q^n$, Theorem 3.2.37 shows that $\mathbf{v} \in C$ if and only if $\mathbf{v}H^\top = \mathbf{0} \in \mathbb{F}_q^{n-k}$. In view of the special form of H, the latter condition means that

$$\sum_{j=0}^{k} f_j v_{j+i} = 0 \qquad \text{for } 0 \leq i \leq n-k-1. \tag{3.13}$$

If $\mathbf{v} = \mathbf{c}(\theta)$, then $v_i = \mathrm{Tr}(\theta\gamma^i)$ for $0 \leq i \leq n-1$, and so

$$\sum_{j=0}^{k} f_j v_{j+i} = \sum_{j=0}^{k} f_j \mathrm{Tr}(\theta\gamma^{j+i}) = \mathrm{Tr}\Big(\sum_{j=0}^{k} f_j \theta\gamma^{j+i}\Big) = \mathrm{Tr}(\theta\gamma^i f(\gamma)) = 0$$

for $0 \leq i \leq n-k-1$ since $f(\gamma) = 0$. Thus, $\mathbf{v} = \mathbf{c}(\theta)$ satisfies the condition (3.13), and so $\mathbf{c}(\theta) \in C$ for all $\theta \in \mathbb{F}_{q^k}$. Since C has exactly q^k codewords, it now suffices to prove that the linear transformation $\theta \in \mathbb{F}_{q^k} \mapsto \mathbf{c}(\theta) \in \mathbb{F}_q^n$ is injective. This boils down to showing that $\mathbf{c}(\theta) = \mathbf{0} \in \mathbb{F}_q^n$ only for $\theta = 0$. If we had $\mathbf{c}(\theta) = \mathbf{0} \in \mathbb{F}_q^n$ for some $\theta \in \mathbb{F}_{q^k}^*$, then $\mathrm{Tr}(\theta\gamma^i) = 0$ for $0 \leq i \leq k-1$. Since $1, \gamma, \gamma^2, \ldots, \gamma^{k-1}$ form a basis of $\mathbb{F}_{q^k}$ over $\mathbb{F}_q$ (compare with Remark 3.2.7), this implies that $\mathrm{Tr}(\beta) = 0$ for all $\beta \in \mathbb{F}_{q^k}$, which is a contradiction to the fact that $\mathrm{Tr} : \mathbb{F}_{q^k} \to \mathbb{F}_q$ is a surjective map by Theorem 1.4.50(iii). □

The explicit formula for the codewords in C given by Theorem 3.3.37, in conjunction with the following simple estimation of character sums, leads to a lower bound and an upper bound on the Hamming weight of each nonzero codeword in C.

Lemma 3.3.38 *Let χ be a nontrivial additive character of the finite field $\mathbb{F}_q$ and let $a \in \mathbb{F}_q^*$ have multiplicative order t. Then*

$$\Big|\sum_{i=0}^{t-1} \chi(ba^i)\Big| \leq (q-t)^{1/2} \qquad \textit{for all } b \in \mathbb{F}_q^*.$$

Proof. We write $s(b)$ for the given character sum and put $s(0) = t$. The sequence $(ba^i)_{i=0}^{\infty}$ is periodic with period length t. Thus, for every integer $j \geq 0$ we obtain

$$s(b) = \sum_{i=0}^{t-1} \chi(ba^{i+j}) = \sum_{i=0}^{t-1} \chi(ba^j a^i) = s(ba^j).$$

The elements $b, ba, ba^2, \ldots, ba^{t-1}$ of $\mathbb{F}_q^*$ are distinct, hence

$$t|s(b)|^2 = \sum_{j=0}^{t-1} |s(ba^j)|^2 \leq \sum_{c \in \mathbb{F}_q^*} |s(c)|^2.$$

Now by expanding $|s(c)|^2$ via $|z|^2 = z\overline{z}$ for all $z \in \mathbb{C}$ and by the orthogonality relation (1.9) for characters, we get

$$\begin{aligned}\sum_{c \in \mathbb{F}_q^*} |s(c)|^2 &= \sum_{c \in \mathbb{F}_q} |s(c)|^2 - |s(0)|^2 = \sum_{c \in \mathbb{F}_q} \sum_{i,j=0}^{t-1} \chi(c(a^i - a^j)) - t^2 \\ &= \sum_{i,j=0}^{t-1} \sum_{c \in \mathbb{F}_q} \chi(c(a^i - a^j)) - t^2 = qt - t^2.\end{aligned}$$

This yields $t|s(b)|^2 \leq qt - t^2$, and so $|s(b)| \leq (q-t)^{1/2}$ as desired. □

Theorem 3.3.39 *If C is an irreducible cyclic $[n,k]$ code over $\mathbb{F}_q$ as in Theorem 3.3.37, then the Hamming weight $w(\mathbf{c})$ of every nonzero codeword $\mathbf{c} \in C$ satisfies*

$$\frac{q-1}{q}\left(n - (q^k - n)^{1/2}\right) \leq w(\mathbf{c}) \leq \frac{q-1}{q}\left(n + (q^k - n)^{1/2}\right).$$

Proof. By Theorem 3.3.37, a nonzero codeword $\mathbf{c} \in C$ is given by $\mathbf{c} = \mathbf{c}(\theta)$ with $\theta \in \mathbb{F}_{q^k}^*$. We write

$$w(\mathbf{c}(\theta)) = n - N(\mathbf{c}(\theta)),$$

where $N(\mathbf{c}(\theta))$ is the number of integers i with $0 \leq i \leq n-1$ and $\mathrm{Tr}(\theta\gamma^i) = 0$. Choose a nontrivial additive character χ of $\mathbb{F}_q$. Then by the orthogonality relation (1.9) for characters we get

$$N(\mathbf{c}(\theta)) = \sum_{i=0}^{n-1} \frac{1}{q} \sum_{b \in \mathbb{F}_q} \chi(b\mathrm{Tr}(\theta\gamma^i)) = \frac{1}{q} \sum_{b \in \mathbb{F}_q} \sum_{i=0}^{n-1} \chi(\mathrm{Tr}(b\theta\gamma^i)).$$

Now $\chi_k(\alpha) = \chi(\mathrm{Tr}(\alpha))$ for $\alpha \in \mathbb{F}_{q^k}$ defines a nontrivial additive character of $\mathbb{F}_{q^k}$, and so we obtain

$$N(\mathbf{c}(\theta)) = \frac{1}{q} \sum_{b \in \mathbb{F}_q} \sum_{i=0}^{n-1} \chi_k(b\theta\gamma^i) = \frac{n}{q} + \frac{1}{q} \sum_{b \in \mathbb{F}_q^*} \sum_{i=0}^{n-1} \chi_k(b\theta\gamma^i).$$

It follows that

$$\left| w(\mathbf{c}(\theta)) - \frac{(q-1)n}{q} \right| \leq \frac{1}{q} \sum_{b \in \mathbb{F}_q^*} \Big| \sum_{i=0}^{n-1} \chi_k(b\theta\gamma^i) \Big|.$$

The last sum is a character sum as in Lemma 3.3.38, with q in that lemma replaced by q^k. By applying the bound in that lemma, we arrive at the desired result. □

Corollary 3.3.40 *If C is an irreducible cyclic $[n,k]$ code over $\mathbb{F}_q$ as in Theorem 3.3.37, then the minimum distance of C satisfies*

$$d(C) \geq \frac{q-1}{q}(n - (q^k - n)^{1/2}).$$

Proof. This follows from Theorems 3.2.14 and 3.3.39. □

The lower bound on $d(C)$ in Corollary 3.3.40 is positive whenever $n > q^{k/2}$. We note that since n is the multiplicative order of the element $\gamma \in \mathbb{F}_{q^k}^*$, the value of n can potentially be as large as $q^k - 1$.

3.3.7 Decoding algorithms for cyclic codes

Since cyclic codes form a special family of linear codes, we can apply the syndrome decoding algorithm (see Algorithm 3.2.52) to cyclic codes. Because of the special structure of cyclic codes, there is some hope that simplifications in this decoding algorithm can be achieved. This is indeed the case if one works with a suitable parity-check matrix of the given cyclic code.

Let $C \subseteq \mathbb{F}_q^n$ be a nontrivial cyclic code with generator polynomial $g(x) \in \mathbb{F}_q[x]$ of degree $n-k$ (note that $1 \le k \le n-1$). Then $\dim(C) = k$ and the syndromes are elements of $\mathbb{F}_q^{n-k}$. We construct a parity-check matrix H of C via its transpose $H^\top$, which is an $n \times (n-k)$ matrix over $\mathbb{F}_q$. First we introduce the linear transformation $\varrho : \mathbb{F}_q[x]_{<n} \to \mathbb{F}_q[x]/(g(x))$ which assigns to each $f(x) \in \mathbb{F}_q[x]_{<n}$ the least residue of $f(x)$ modulo $g(x)$, that is, the remainder of $f(x)$ after division by $g(x)$. Then we set up the vector space isomorphism $\tau : \mathbb{F}_q[x]/(g(x)) \to \mathbb{F}_q^{n-k}$ which sends each least residue modulo $g(x)$ (which is a polynomial of degree less than $n-k$) to its coefficient vector, in analogy with the inverse of the map π in (3.9). The composite map $\tau \circ \varrho : \mathbb{F}_q[x]_{<n} \to \mathbb{F}_q^{n-k}$ is again a linear transformation between vector spaces over $\mathbb{F}_q$.

Now we construct the matrix $H^\top$ by letting its jth row be $(\tau \circ \varrho)(x^{j-1})$ for $1 \le j \le n$. For $\mathbf{v} \in \mathbb{F}_q^n$ and its corresponding polynomial $v = \pi(\mathbf{v}) \in \mathbb{F}_q[x]_{<n}$, we obtain the logical equivalences

$$\mathbf{v} \in C \Leftrightarrow \varrho(v) = 0 \in \mathbb{F}_q[x]/(g(x)) \Leftrightarrow (\tau \circ \varrho)(v) = \mathbf{0} \in \mathbb{F}_q^{n-k} \Leftrightarrow \mathbf{v}H^\top = \mathbf{0} \in \mathbb{F}_q^{n-k}.$$

Furthermore, the first $n-k$ rows of $H^\top$ form the identity matrix I_{n-k}, and so the column vectors of $H^\top$ are linearly independent over $\mathbb{F}_q$. Consequently, H is a parity-check matrix of C.

It is convenient to carry out the syndrome decoding algorithm for cyclic codes in the language of polynomials. To this end, we translate syndromes in $\mathbb{F}_q^{n-k}$ into polynomials, by applying the inverse $\tau^{-1} : \mathbb{F}_q^{n-k} \to \mathbb{F}_q[x]/(g(x))$ of the vector space isomorphism τ introduced above. For every $\mathbf{v} = (v_0, v_1, \ldots, v_{n-1}) \in \mathbb{F}_q^n$ and its syndrome $S(\mathbf{v}) = \mathbf{v}H^\top$, an application of τ^{-1} to $S(\mathbf{v})$ amounts to multiplying $\mathbf{v}$ by the $n \times (n-k)$ matrix P whose jth row is $\varrho(x^{j-1})$ for $1 \le j \le n$. Therefore

$$\tau^{-1}(S(\mathbf{v})) = \mathbf{v}P = \sum_{j=0}^{n-1} v_j \varrho(x^j) = \varrho\Big(\sum_{j=0}^{n-1} v_j x^j\Big) = \varrho(v),$$

where $v = v(x) = \sum_{j=0}^{n-1} v_j x^j \in \mathbb{F}_q[x]_{<n}$ is the polynomial corresponding to $\mathbf{v}$. Thus, in the context of the syndrome decoding algorithm for cyclic codes, it is convenient to speak of the *received polynomial* $v(x) \in \mathbb{F}_q[x]_{<n}$ instead of the received word $\mathbf{v} \in \mathbb{F}_q^n$, and the corresponding syndrome can then be viewed as $\tau^{-1}(S(\mathbf{v})) = \varrho(v)$, that is, the least residue of $v(x)$ modulo $g(x)$. As in Subsection 3.3.1, we identify

$\mathbb{F}_q[x]_{<n}$ with the residue class ring $\mathbb{F}_q[x]/(x^n-1)$. We summarize this in the following definition.

Definition 3.3.41 Let $C \subseteq \mathbb{F}_q^n$ be a nontrivial cyclic code with generator polynomial $g(x)$. Then for a received polynomial $v = v(x) \in \mathbb{F}_q[x]/(x^n - 1)$, its *syndrome polynomial* $S(v) \in \mathbb{F}_q[x]/(g(x))$ is the least residue of $v(x)$ modulo $g(x)$.

Note that $S(v)$ is well defined for $v \in \mathbb{F}_q[x]/(x^n-1)$ because $g(x)$ divides x^n-1 in $\mathbb{F}_q[x]$ by Theorem 3.3.12. Since the syndrome $S(\mathbf{v})$ in the sense of Definition 3.2.49 and the syndrome polynomial $S(v)$ in the sense of Definition 3.3.41 correspond to each other via a vector space isomorphism, it is clear that $S(v)$ shares the properties of $S(\mathbf{v})$ in Proposition 3.2.51.

For consistency, we interpret an error word $\mathbf{e} = (e_0, e_1, \ldots, e_{n-1}) \in \mathbb{F}_q^n$ also as a polynomial, namely as the *error polynomial* $e(x) = \sum_{j=0}^{n-1} e_j x^j \in \mathbb{F}_q[x]$. From (3.7) and with a received polynomial $v(x)$, we then obtain the *code polynomial* $c(x) = v(x) - e(x)$ which belongs to the ideal of $\mathbb{F}_q[x]/(x^n - 1)$ consisting exactly of all multiples of the generator polynomial $g(x)$. The following is an easy situation in which a most likely error polynomial can be obtained immediately.

Proposition 3.3.42 *Let C be a nontrivial cyclic $[n,k,d]$ code over $\mathbb{F}_q$. If for a received polynomial $v = v(x) \in \mathbb{F}_q[x]/(x^n - 1)$ the syndrome polynomial $S(v)$ has at most $\lfloor (d-1)/2 \rfloor$ nonzero coefficients, then $S(v)$ is the most likely error polynomial.*

Proof. By Definition 3.3.41, $S(v) - v$ is a multiple of the generator polynomial $g(x)$ of C and so a code polynomial. In other words, $S(v)$ and v are in the same coset of C. We are done if we can prove that $S(v)$ is the unique coset leader of this coset. Suppose that there is another polynomial $w \in \mathbb{F}_q[x]/(x^n - 1)$ in this coset with at most $\lfloor (d-1)/2 \rfloor$ nonzero coefficients. Then $S(v) - w$ is a code polynomial. But $S(v) - w$ has $\leq 2\lfloor (d-1)/2 \rfloor \leq d-1$ nonzero coefficients, that is, it corresponds to a codeword $\mathbf{c} \in C$ of Hamming weight at most $d-1$. Since C has minimum distance d, it follows that $\mathbf{c} = \mathbf{0} \in \mathbb{F}_q^n$, hence $w = S(v)$. □

Example 3.3.43 Let C be the binary cyclic code of length 7 in Example 3.3.22. From the generator matrix G in Example 3.3.22 we can easily determine all codewords in C and check that $d(C) = 3$. Suppose that the received polynomial is $v(x) = x + x^2 + x^4 + x^5$. Dividing $v(x)$ by the generator polynomial $g(x) = 1 + x^2 + x^3 \in \mathbb{F}_2[x]$, we get the quotient x^2 and the remainder x. Therefore $S(v) = x$. This syndrome polynomial satisfies the condition in Proposition 3.3.42, and so the most likely error polynomial is $e(x) = x$. The most likely sent code polynomial is $c(x) = v(x) - e(x) = x^2 + x^4 + x^5$, which corresponds to the codeword $(0\ 0\ 1\ 0\ 1\ 1\ 0)$ in C.

In preparation for a more refined version of the syndrome decoding algorithm for cyclic codes, we study how the syndrome polynomial changes under a cyclic shift of the input.

Lemma 3.3.44 *Let C be a nontrivial cyclic $[n, k]$ code over $\mathbb{F}_q$ with generator polynomial $g(x)$. Let $S(v) = S(v(x))$ be the syndrome polynomial of a received polynomial $v = v(x) \in \mathbb{F}_q[x]/(x^n - 1)$. Then*

$$S(xv(x)) = xS(v(x)) - s_{n-k-1}g(x),$$

where s_{n-k-1} is the coefficient of x^{n-k-1} in the polynomial $S(v)$.

Proof. By the definition of $S(v(x))$, we can write $v(x) = a(x)g(x) + S(v(x))$ for some $a(x) \in \mathbb{F}_q[x]$, where $\deg(S(v(x))) < \deg(g(x)) = n - k$. Then

$$xv(x) = xa(x)g(x) + xS(v(x)) = (xa(x) + s_{n-k-1})g(x) + xS(v(x)) - s_{n-k-1}g(x).$$

Furthermore,

$$\deg(xS(v(x)) - s_{n-k-1}g(x)) \leq \max\left(\deg(xS(v(x))), \deg(s_{n-k-1}g(x))\right) \leq n - k.$$

The coefficient of x^{n-k} in $xS(v(x)) - s_{n-k-1}g(x)$ is $s_{n-k-1} - s_{n-k-1} \cdot 1 = 0$ since $g(x)$ is a monic polynomial. Therefore

$$\deg(xS(v(x)) - s_{n-k-1}g(x)) < n - k,$$

and so $xS(v(x)) - s_{n-k-1}g(x)$ is the least residue of $xv(x)$ modulo $g(x)$. □

Remark 3.3.45 Given the syndrome polynomial of the cyclic shift $x^t v(x)$ of a received polynomial $v(x) \in \mathbb{F}_q[x]/(x^n - 1)$, the syndrome polynomial of $x^{t+1}v(x)$ can be computed by means of Lemma 3.3.44. Thus, the syndrome polynomials of $xv(x), x^2v(x), \ldots$ can be computed recursively.

Definition 3.3.46 A word $\mathbf{u} = (u_0, u_1, \ldots, u_{n-1}) \in \mathbb{F}_q^n$ has a *cyclic run of zeros* of length $\ell \geq 1$ if it has a succession of ℓ cyclically consecutive zero coordinates.

Example 3.3.47 The word $\mathbf{u} = (0, 0, 1, 0, 0, 0, 0, 1, 0, 0) \in \mathbb{F}_2^9$ has a cyclic run of zeros of length 4.

Here is another auxiliary result that we need for a refined syndrome decoding algorithm for cyclic codes.

Lemma 3.3.48 *Let C be a nontrivial cyclic $[n,k]$ code over $\mathbb{F}_q$. Suppose that for some received polynomial $v(x) \in \mathbb{F}_q[x]/(x^n-1)$ we have an error word $\mathbf{e} \in \mathbb{F}_q^n$ which has a cyclic run of zeros of length at least k. Then there exists an integer $t \geq 0$ such that for $h = h(x) = x^t v(x) \in \mathbb{F}_q[x]/(x^n-1)$ the syndrome polynomial is given by $S(h) = r(x)$, where $r(x) \in \mathbb{F}_q[x]/(x^n-1)$ is the polynomial corresponding to the cyclic shift $\mathbf{e}^t$. Furthermore, the number of nonzero coefficients of $r(x)$ is equal to the Hamming weight $w(\mathbf{e})$.*

Proof. Since $\mathbf{e}$ has a cyclic run of zeros of length at least k, there exists an integer $t \geq 0$ such that the last k coordinates of the cyclic shift $\mathbf{e}^t$ are all equal to 0. Thus, if $r(x) \in \mathbb{F}_q[x]/(x^n-1)$ is as in the lemma, then $\deg(r(x)) < n-k$. Since $\deg(g(x)) = n-k$, it follows from Definition 3.3.41 that $S(r(x)) = r(x)$. If $e(x) \in \mathbb{F}_q[x]/(x^n-1)$ is the polynomial corresponding to $\mathbf{e}$, then

$$h(x) \equiv x^t v(x) \equiv x^t e(x) \equiv r(x) \pmod{g(x)},$$

and so $S(h) = S(r(x)) = r(x)$. The last part of the lemma is trivial since a cyclic shift does not change the Hamming weight of a word. □

We are now ready to describe a syndrome decoding algorithm for cyclic codes which is applicable under less restrictive conditions than those in Proposition 3.3.42. This algorithm goes by a colorful name suggesting that we want to capture the devious error in a trap.

Algorithm 3.3.49 (Error-Trapping Decoding Algorithm for Cyclic Codes) Let C be a nontrivial cyclic $[n,k,d]$ code over $\mathbb{F}_q$. For a received word $\mathbf{v} \in \mathbb{F}_q^n$, suppose that the error word $\mathbf{e} \in \mathbb{F}_q^n$ has Hamming weight $w(\mathbf{e}) \leq \lfloor (d-1)/2 \rfloor$ and a cyclic run of zeros of length at least k.
Step 1: For the received polynomial $v(x) \in \mathbb{F}_q[x]/(x^n-1)$ corresponding to $\mathbf{v}$, compute the syndrome polynomials of $v(x), xv(x), x^2v(x), \ldots$ by Lemma 3.3.44 until an integer t with $0 \leq t \leq n-1$ is obtained for which the syndrome polynomial of $x^t v(x) \in \mathbb{F}_q[x]/(x^n-1)$ has at most $\lfloor (d-1)/2 \rfloor$ nonzero coefficients. Such an integer t exists by Lemma 3.3.48.
Step 2: By Proposition 3.3.42, $e_t(x) := S(x^t v(x))$ is the most likely error polynomial for the received polynomial $x^t v(x)$. Let $\mathbf{e}_t \in \mathbb{F}_q^n$ be the word corresponding to $e_t(x)$. A cyclic shift of $\mathbf{e}_t$ by $n-t$ positions yields the most likely error word $\mathbf{e}$.
Step 3: Compute the most likely sent codeword $\mathbf{c}$ as $\mathbf{c} = \mathbf{v} - \mathbf{e}$.

Example 3.3.50 Let C be the cyclic $[7,4,3]$ code over $\mathbb{F}_2$ with generator polynomial $g(x) = x^3 + x^2 + 1$ considered in Examples 3.3.22 and 3.3.43. Let the received polynomial be $v(x) = 1 + x^2 + x^3 + x^4$. As in the error-trapping decoding algorithm above, we assume that the error word $\mathbf{e} \in \mathbb{F}_2^7$ has Hamming weight $w(\mathbf{e}) \leq 1$. Then

$\mathbf{e}$ must have a cyclic run of zeros of length at least 6, and so all conditions in the algorithm above are satisfied. Dividing $v(x)$ by $g(x)$, we get

$$v(x) = xg(x) + 1 + x + x^2,$$

and so $S(v(x)) = 1 + x + x^2$. The condition in Step 1 of the algorithm is thus not satisfied for $t = 0$. Therefore we proceed to $t = 1$. According to Lemma 3.3.44, we obtain

$$S(xv(x)) = x(1 + x + x^2) - 1 \cdot (1 + x^2 + x^3) = 1 + x.$$

The condition in Step 1 of the algorithm is thus not satisfied for $t = 1$. Therefore we proceed to $t = 2$. Again by Lemma 3.3.44, we compute

$$S(x^2v(x)) = x(1 + x) - 0 \cdot (1 + x^2 + x^3) = x + x^2.$$

The condition in Step 1 of the algorithm is thus not satisfied for $t = 2$. But we are not discouraged since we know from Lemma 3.3.48 that a suitable value of t must exist. Hence we proceed to $t = 3$. Again by Lemma 3.3.44, we obtain

$$S(x^3v(x)) = x(x + x^2) - 1 \cdot (1 + x^2 + x^3) = 1.$$

Now the condition in Step 1 of the algorithm is satisfied. In Step 2 of the algorithm we get $e_3(x) = 1$ and $\mathbf{e}_3 = (1\ 0\ 0\ 0\ 0\ 0\ 0) \in \mathbb{F}_2^7$. A cyclic shift of $\mathbf{e}_3$ by $n-t = 7-3 = 4$ positions yields $\mathbf{e} = (0\ 0\ 0\ 0\ 1\ 0\ 0) \in \mathbb{F}_2^7$. The most likely sent codeword is therefore

$$\mathbf{c} = (1\ 0\ 1\ 1\ 1\ 0\ 0) - (0\ 0\ 0\ 0\ 1\ 0\ 0) = (1\ 0\ 1\ 1\ 0\ 0\ 0).$$

A transmission error occurred most likely in the fifth coordinate.

3.4 Bounds in coding theory

3.4.1 Existence theorems for good codes

There are some general theoretical results that establish the existence of good codes provided that the parameters satisfy certain bounds. As usual, we write $|A|$ for the cardinality (that is, the number of elements) of a finite set A.

Theorem 3.4.1 (Sphere-Covering Bound) *If $\mathbb{F}_q$ is a finite field and n and d are integers with $1 \leq d \leq n$, then there exists a code $C \subseteq \mathbb{F}_q^n$ with minimum distance d and*

$$|C| \sum_{i=0}^{d-1} \binom{n}{i} (q-1)^i \geq q^n.$$

Proof. For fixed n, d, and q, let $C \subseteq \mathbb{F}_q^n$ be a code of maximum size with $d(C) = d$. If $\mathbf{w} \in \mathbb{F}_q^n$ were such that $d(\mathbf{c}, \mathbf{w}) \geq d$ for all $\mathbf{c} \in C$, then the larger code $C \cup \{\mathbf{w}\}$ would still have minimum distance d. Thus, there can be no such $\mathbf{w} \in \mathbb{F}_q^n$. In other words, for every $\mathbf{v} \in \mathbb{F}_q^n$ there is a $\mathbf{c} \in C$ with $d(\mathbf{c}, \mathbf{v}) \leq d - 1$. Hence

$$\mathbb{F}_q^n = \bigcup_{\mathbf{c} \in C} B(\mathbf{c}, d-1),$$

where $B(\mathbf{c}, d-1)$ is the ball with center $\mathbf{c}$ and radius $d-1$ in the Hamming space $\mathbb{F}_q^n$ defined by

$$B(\mathbf{c}, d-1) = \{\mathbf{v} \in \mathbb{F}_q^n : d(\mathbf{c}, \mathbf{v}) \leq d-1\}.$$

By considering cardinalities, we get

$$q^n = \Big| \bigcup_{\mathbf{c} \in C} B(\mathbf{c}, d-1) \Big| \leq \sum_{\mathbf{c} \in C} |B(\mathbf{c}, d-1)| = |C| \sum_{i=0}^{d-1} \binom{n}{i} (q-1)^i,$$

which is the desired result. Note that the formula for $|B(\mathbf{c}, d-1)|$ above is obtained by first fixing the number i of coordinates where $\mathbf{c}$ and $\mathbf{v}$ differ (with $0 \leq i \leq d-1$), then choosing the $\binom{n}{i}$ actual coordinate positions where $\mathbf{c}$ and $\mathbf{v}$ differ, and finally observing that for i fixed coordinate positions this leaves exactly $(q-1)^i$ possibilities for $\mathbf{v}$. □

Example 3.4.2 Let $q = 2$, $n = 9$, and $d = 3$. Then Theorem 3.4.1 ensures the existence of a binary code C of length 9 with minimum distance 3 and $|C| \geq 512/46$, that is, with $|C| \geq 12$.

The proof of Theorem 3.4.1 provides no guarantee that the code in this theorem is linear. However, there is the following result of roughly comparable quality for linear codes.

Theorem 3.4.3 (Gilbert-Varshamov Bound) *Let n, k, and d be integers with $1 \leq k < n$, $2 \leq d \leq n$, and*

$$\sum_{i=0}^{d-2} \binom{n-1}{i} (q-1)^i < q^{n-k}. \tag{3.14}$$

Then there exists a linear $[n, k]$ code over $\mathbb{F}_q$ with minimum distance at least d.

Proof. We first observe that

$$q^{d-2} = \sum_{i=0}^{d-2} \binom{d-2}{i} (q-1)^i \leq \sum_{i=0}^{d-2} \binom{n-1}{i} (q-1)^i < q^{n-k}$$

by (3.14), and so $d - 1 \leq n - k$. Now we construct a certain $(n-k) \times n$ matrix H' over $\mathbb{F}_q$ columnwise. We choose the first $d-1$ column vectors of H' as linearly independent vectors from $\mathbb{F}_q^{n-k}$ (this is possible since $d - 1 \leq n - k$). Now suppose that the first $j-1$ column vectors of H' (with $d \leq j \leq n$) have already been constructed and satisfy the property that any $d-1$ of them are linearly independent over $\mathbb{F}_q$. There are at most

$$\sum_{i=0}^{d-2} \binom{j-1}{i} (q-1)^i \leq \sum_{i=0}^{d-2} \binom{n-1}{i} (q-1)^i$$

vectors from $\mathbb{F}_q^{n-k}$ that can be obtained as linear combinations over $\mathbb{F}_q$ of $d-2$ or fewer of these $j-1$ column vectors. Since (3.14) holds, it is possible to choose a jth column vector of H' that is linearly independent of any $d-2$ of the first $j-1$ column vectors of H'. When this inductive construction is complete, we arrive at an $(n-k) \times n$ matrix H' over $\mathbb{F}_q$ with the property that any $d-1$ column vectors of H' are linearly independent over $\mathbb{F}_q$. The null space of H' (see Remark 3.2.39) is a linear code C' over $\mathbb{F}_q$ of length n with $\dim(C') \geq k$. Furthermore, the argument in the proof of Theorem 3.2.44 shows that $d(C') \geq d$. Now let C be an arbitrary k-dimensional subspace of C'. Since passing to a subspace cannot decrease the minimum distance, we see that C is a linear code of the desired type. □

Example 3.4.4 Let $q = 2$, $n = 7$, $k = 4$, and $d = 3$. Then the inequality (3.14) is satisfied, and so Theorem 3.4.3 shows the existence of a binary linear $[7, 4]$ code with minimum distance at least 3. A simple explicit construction of such a code was given in Example 3.2.53.

Example 3.4.5 Let $q = 3$, $n = 10$, $k = 7$, and $d = 3$. Then the inequality (3.14) is satisfied, and so Theorem 3.4.3 guarantees the existence of a linear $[10, 7]$ code over $\mathbb{F}_3$ with minimum distance at least 3.

The procedure in the proof of Theorem 3.4.3 can, in principle, be implemented to construct good linear codes. However, it should be noted that, for large values of d, this method is usually impracticable. We will present efficient explicit constructions of families of good linear codes later in this chapter.

3.4.2 Limitations on the parameters of codes

The parameters of codes cannot be chosen independently of each other. For instance, it is obvious that for a linear $[n, k, d]$ code the bounds $1 \leq k \leq n$ and $1 \leq d \leq n$ are valid. In the following, we will discuss less trivial limitations on the parameters of codes.

Theorem 3.4.6 (Hamming Bound) *Every t-error-correcting code $C \subseteq \mathbb{F}_q^n$ satisfies*

$$|C| \sum_{i=0}^{t} \binom{n}{i} (q-1)^i \le q^n.$$

Proof. For all $\mathbf{c} \in C$, let $B(\mathbf{c}, t) \subseteq \mathbb{F}_q^n$ be the ball with center $\mathbf{c}$ and radius t (compare with the proof of Theorem 3.4.1). For distinct $\mathbf{c}_1, \mathbf{c}_2 \in C$, it follows from Definition 3.1.12 that $B(\mathbf{c}_1, t)$ and $B(\mathbf{c}_2, t)$ are disjoint. Thus, by comparing cardinalities in

$$\bigcup_{\mathbf{c} \in C} B(\mathbf{c}, t) \subseteq \mathbb{F}_q^n$$

and referring again to the proof of Theorem 3.4.1, we obtain the desired inequality. □

Corollary 3.4.7 *Every code $C \subseteq \mathbb{F}_q^n$ with $|C| \ge 2$ and minimum distance d satisfies*

$$|C| \sum_{i=0}^{\lfloor (d-1)/2 \rfloor} \binom{n}{i} (q-1)^i \le q^n. \tag{3.15}$$

Proof. This follows from Theorems 3.1.14 and 3.4.6. □

Definition 3.4.8 A code $C \subseteq \mathbb{F}_q^n$ with $|C| \ge 2$ and minimum distance d which achieves equality in (3.15) is called *perfect*.

Example 3.4.9 The trivial code $C = \mathbb{F}_q^n$ is obviously perfect. For $q = 2$, the repetition code of odd length n (see Example 3.1.4) is easily seen to be perfect. A simple computation shows that the binary linear $[7, 4, 3]$ code constructed in Example 3.2.53 is perfect. The last example will be generalized in Theorem 3.5.7. Further examples of perfect codes will be presented in Subsection 3.5.2.

An inspection of the proof of Theorem 3.4.6 reveals that a perfect code $C \subseteq \mathbb{F}_q^n$ has the intriguing geometric property that the balls of radius $\lfloor (d(C)-1)/2 \rfloor$ around the codewords in C are disjoint and fill up the whole space $\mathbb{F}_q^n$. This is like tightly packing oranges in $\mathbb{F}_q^n$ as if they were cubes; in $\mathbb{R}^n$ we can do this only with orange juice. Because of this interpretation in terms of packing, the bound in Theorem 3.4.6 is called also the *sphere-packing bound*.

The following bound provides another important restriction on the parameters of a code. The name of this bound has nothing to do with "singleton" in the sense of a one-element set, but rather with the coding theorist Richard Singleton and his paper [188].

Theorem 3.4.10 (Singleton Bound) *Every code $C \subseteq \mathbb{F}_q^n$ with $|C| \geq 2$ and minimum distance d satisfies*

$$|C| \leq q^{n-d+1}.$$

Proof. In each codeword in C we delete the last $d-1$ coordinates. The resulting words over $\mathbb{F}_q$ of length $n-d+1$ are distinct since C has minimum distance d. Therefore $|C| \leq q^{n-d+1}$, the total number of words over $\mathbb{F}_q$ of length $n-d+1$. □

Corollary 3.4.11 (Singleton Bound for Linear Codes) *Every linear $[n,k,d]$ code over $\mathbb{F}_q$ satisfies*

$$d \leq n-k+1.$$

Proof. This follows from Theorem 3.4.10 since a linear $[n,k,d]$ code C over $\mathbb{F}_q$ satisfies $|C| = q^k$. □

Remark 3.4.12 There is a simple alternative proof of Corollary 3.4.11 in which we turn to an equivalent linear code with a generator matrix G' in standard form (see Remark 3.2.27) and then note that all row vectors of G' have Hamming weight at most $n-k+1$.

Theorem 3.4.10 substantiates a remark we made at the end of Subsection 3.1.2, namely that there is a trade-off between the desiderata of a large minimum distance of a code $C \subseteq \mathbb{F}_q^n$ and a large data transmission rate of C (which can be expressed by saying that the ratio $|C|/q^n$ is relatively large). Indeed, Theorem 3.4.10 demonstrates that if the minimum distance d of C is large, then the ratio $|C|/q^n$ is necessarily small.

Definition 3.4.13 A linear $[n,k,d]$ code over $\mathbb{F}_q$ with $d = n-k+1$ is called an *MDS code.*

The acronym MDS stands for *maximum distance separable*, which is suggestive of the fact that an MDS code is a linear $[n,k]$ code that achieves the largest minimum distance $n-k+1$ allowed by the Singleton bound for linear codes. After giving some easy examples of MDS codes, we discuss basic properties of MDS codes.

Example 3.4.14 The trivial code $C_1 = \mathbb{F}_q^n$ and the linear $[n,1,n]$ code C_2 over $\mathbb{F}_q$ with basis vector $\mathbf{b} = (b_1, \ldots, b_n) \in \mathbb{F}_q^n$, where $b_j \neq 0$ for $1 \leq j \leq n$, are MDS codes. For every $n \geq 2$, the linear $[n, n-1, 2]$ code C_3 over $\mathbb{F}_q$ with basis vectors $\mathbf{b}_1, \ldots, \mathbf{b}_{n-1} \in \mathbb{F}_q^n$ is an MDS code, where for $i = 1, \ldots, n-1$ the vector $\mathbf{b}_i$ has coordinate 1 in positions i and $i+1$ and coordinate 0 elsewhere.

Proposition 3.4.15 *Let C be a nontrivial linear $[n, k]$ code over $\mathbb{F}_q$ with parity-check matrix H. Then C is an MDS code if and only if any $n - k$ column vectors of H are linearly independent over $\mathbb{F}_q$.*

Proof. In view of Corollary 3.4.11, C is an MDS code if and only if $d(C) \geq n-k+1$. The rest follows from Theorem 3.2.44. □

Theorem 3.4.16 *If the nontrivial linear code C is an MDS code, then its dual code $C^\perp$ is also an MDS code.*

Proof. Let C be a nontrivial linear $[n, k]$ code over $\mathbb{F}_q$ which is an MDS code and let H be a parity-check matrix of C. Then $C^\perp$ is a linear $[n, n-k]$ code over $\mathbb{F}_q$ by Theorem 3.2.34. If $d = d(C^\perp)$, then $d \leq k+1$ by Corollary 3.4.11. In order to make sure that $C^\perp$ is an MDS code, it therefore suffices to prove that the minimum Hamming weight $w(C^\perp)$ is at least $k+1$ (compare with Theorem 3.2.14). Take a codeword $\mathbf{u} = (u_1, \ldots, u_n) \in C^\perp$ with $w(\mathbf{u}) \leq k$. Then there exist integers $1 \leq j_1 < j_2 < \cdots < j_{n-k} \leq n$ with

$$u_{j_1} = u_{j_2} = \cdots = u_{j_{n-k}} = 0. \tag{3.16}$$

Since H is a generator matrix of $C^\perp$, there is some $\mathbf{a} \in \mathbb{F}_q^{n-k}$ for which $\mathbf{u} = \mathbf{a}H$. If $\mathbf{h}_1^\top, \ldots, \mathbf{h}_n^\top$ are the column vectors of H, then (3.16) implies that $\mathbf{a} \cdot \mathbf{h}_{j_1} = \mathbf{a} \cdot \mathbf{h}_{j_2} = \cdots = \mathbf{a} \cdot \mathbf{h}_{j_{n-k}} = 0$. Now $\mathbf{h}_{j_1}, \mathbf{h}_{j_2}, \ldots, \mathbf{h}_{j_{n-k}}$ are linearly independent over $\mathbb{F}_q$ by Proposition 3.4.15, and so these vectors form a basis of $\mathbb{F}_q^{n-k}$. It follows that $\mathbf{a} \cdot \mathbf{v} = 0$ for all $\mathbf{v} \in \mathbb{F}_q^{n-k}$, hence $\mathbf{a} = \mathbf{0} \in \mathbb{F}_q^{n-k}$ and $\mathbf{u} = \mathbf{0} \in \mathbb{F}_q^n$. This shows that $w(C^\perp) \geq k+1$. □

Proposition 3.4.17 *Let C be a linear $[n, k]$ code over $\mathbb{F}_q$ with generator matrix G. Then C is an MDS code if and only if any k column vectors of G are linearly independent over $\mathbb{F}_q$.*

Proof. This is trivial for $k = n$. If $1 \leq k \leq n-1$, then Proposition 3.4.15 shows that any k column vectors of G are linearly independent over $\mathbb{F}_q$ if and only if $C^\perp$ is an MDS code (since G is a parity-check matrix of $C^\perp$ and $\dim(C^\perp) = n-k$). Now $(C^\perp)^\perp = C$ by Corollary 3.2.35, and so we deduce from Theorem 3.4.16 that $C^\perp$ is an MDS code if and only if C is an MDS code. □

Example 3.4.18 Let C be the linear $[4, 2]$ code over $\mathbb{F}_3$ with generator matrix

$$G = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 2 \end{pmatrix}.$$

Then any two column vectors of G are linearly independent over $\mathbb{F}_3$, and so C is an MDS code by Proposition 3.4.17.

An important family of MDS codes, namely that of Reed-Solomon codes, will be introduced in Subsection 3.5.3. Now we present another bound on code parameters. Since our focus in this chapter is on linear codes, we establish the bound here only for these codes. There is an analogous bound for arbitrary codes (see [106, Section 5.5] and Theorem 6.4.13).

Theorem 3.4.19 (Plotkin Bound for Linear Codes) *Every linear $[n,k,d]$ code over $\mathbb{F}_q$ satisfies*

$$d \leq \frac{n(q-1)q^{k-1}}{q^k-1}.$$

Proof. For a given linear $[n,k,d]$ code C over $\mathbb{F}_q$ and for $j=1,\ldots,n$, let λ_j be the linear transformation

$$\lambda_j : (c_1,\ldots,c_n) \in C \mapsto c_j \in \mathbb{F}_q.$$

If w_1 denotes the Hamming weight of words over $\mathbb{F}_q$ of length 1, then

$$\sum_{\mathbf{c}\in C} w(\mathbf{c}) = \sum_{\mathbf{c}\in C}\sum_{j=1}^{n} w_1(\lambda_j(\mathbf{c})) = \sum_{j=1}^{n}\sum_{\mathbf{c}\in C} w_1(\lambda_j(\mathbf{c})).$$

If the image of λ_j is $\{0\}$, then the last inner sum is equal to 0. Otherwise, for every $b \in \mathbb{F}_q$ there are exactly q^{k-1} codewords $\mathbf{c} \in C$ with $\lambda_j(\mathbf{c}) = b$, and so the last inner sum is equal to $(q-1)q^{k-1}$. Altogether, we get

$$\sum_{\mathbf{c}\in C} w(\mathbf{c}) \leq n(q-1)q^{k-1}.$$

On the other hand, it is trivial that

$$\sum_{\mathbf{c}\in C} w(\mathbf{c}) \geq (q^k-1)d,$$

and so the desired bound follows. □

Remark 3.4.20 The ternary linear $[4,2,3]$ code in Example 3.4.18 is not only an MDS code, but it also achieves equality in the Plotkin bound. In general, if a linear $[n,k,d]$ code C over $\mathbb{F}_q$ achieves equality in the Plotkin bound, then $\sum_{\mathbf{c}\in C} w(\mathbf{c}) = (q^k-1)d$, as we see from the proof of Theorem 3.4.19. This means that $w(\mathbf{c}) = d$ for all nonzero codewords $\mathbf{c} \in C$. Since $d(\mathbf{u},\mathbf{v}) = w(\mathbf{u}-\mathbf{v})$ for all $\mathbf{u},\mathbf{v} \in \mathbb{F}_q^n$ by (3.5), it follows that $d(\mathbf{c}_1,\mathbf{c}_2) = d$ for any two distinct codewords $\mathbf{c}_1,\mathbf{c}_2 \in C$. For obvious reasons, such a code is called an *equidistant code.*

Example 3.4.21 Given a finite field $\mathbb{F}_q$ and an integer k with $q^k \geq 3$, let $f(x)$ be the minimal polynomial of a primitive element $\gamma \in \mathbb{F}_{q^k}$ over $\mathbb{F}_q$. Let C be the irreducible cyclic $[q^k - 1, k]$ code over $\mathbb{F}_q$ with parity-check polynomial $f(x)$ (see Theorem 3.3.37). Every nonzero $\mathbf{c} \in C$ has the form $\mathbf{c}(\theta)$ in Theorem 3.3.37 for some $\theta \in \mathbb{F}_{q^k}^*$. It follows that $w(\mathbf{c})$ is equal to the number of $\beta \in \mathbb{F}_{q^k}^*$ with $\mathrm{Tr}(\beta) \neq 0$. Since $\mathrm{Tr}(0) = 0$ and the map $\mathrm{Tr} : \mathbb{F}_{q^k} \to \mathbb{F}_q$ attains each value in $\mathbb{F}_q$ equally often (namely q^{k-1} times by Theorem 1.4.50(iii)), we deduce that $w(\mathbf{c}) = (q-1)q^{k-1}$ for every nonzero $\mathbf{c} \in C$. Therefore C is an equidistant code and it achieves equality in the Plotkin bound.

3.5 Some special linear codes

3.5.1 Hamming codes

For every finite field $\mathbb{F}_q$, there is an infinite family of perfect linear codes over $\mathbb{F}_q$, namely that of Hamming codes over $\mathbb{F}_q$. These codes are named after the work of Hamming in his article [61], which is one of the early fundamental papers on coding theory, and they are obtained by an elegant construction. For ease of explanation, we commence with the simpler binary case.

For an integer $r \geq 2$, consider an $r \times (2^r - 1)$ matrix H_r over $\mathbb{F}_2$ whose column vectors are exactly all $2^r - 1$ nonzero vectors from $\mathbb{F}_2^r$. For instance, for $r = 3$ we can take

$$H_3 = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

Here we have arranged the columns in lexicographic order, with the rule that 0 precedes 1, but this is not necessary. Another possible matrix H_r for $r = 3$ is the matrix H in Example 3.2.53. For every $r \geq 2$, we find among the column vectors of H_r in particular all vectors of Hamming weight 1, and so the row vectors of H_r are linearly independent over $\mathbb{F}_2$. Therefore H_r can be chosen as a parity-check matrix of a binary linear code. The length of the resulting code is $n = 2^r - 1$ and its dimension is $n - r = 2^r - 1 - r$.

Definition 3.5.1 For an integer $r \geq 2$, let H_r be an $r \times (2^r - 1)$ matrix over $\mathbb{F}_2$ whose column vectors are exactly all $2^r - 1$ nonzero vectors from $\mathbb{F}_2^r$. Then the linear $[2^r - 1, 2^r - 1 - r]$ code over $\mathbb{F}_2$ with parity-check matrix H_r is called a *binary Hamming code* $\mathrm{Ham}(r, 2)$.

The order of the columns of H_r has not been fixed, and so $\mathrm{Ham}(r, 2)$ is well defined only up to equivalence of codes (see Remark 3.2.27 for the latter notion). Therefore we speak of *a* binary Hamming code $\mathrm{Ham}(r, 2)$ and not of *the* binary Hamming code $\mathrm{Ham}(r, 2)$, and similarly for related codes. The minimum distance of $\mathrm{Ham}(r, 2)$ can be easily determined by Corollary 3.2.45. Note that all column vectors of H_r are nonzero and different, and so any two column vectors of H_r are linearly independent over $\mathbb{F}_2$. On the other hand, the three column vectors $(1\ 0\ 0\ \dots\ 0)^\top$, $(0\ 1\ 0\ \dots\ 0)^\top$, and $(1\ 1\ 0\ \dots\ 0)^\top$ of H_r are linearly dependent over $\mathbb{F}_2$, and so $\mathrm{Ham}(r, 2)$ has minimum distance 3.

Remark 3.5.2 For every integer $r \geq 2$, a suitable code $\mathrm{Ham}(r, 2)$ in the equivalence class is cyclic with generator polynomial $g(x) \in \mathbb{F}_2[x]$, where $g(x)$ is the minimal polynomial of a primitive element $\alpha \in \mathbb{F}_{2^r}$ over $\mathbb{F}_2$. In order to prove this claim, we set up the $r \times (2^r - 1)$ matrix H over $\mathbb{F}_2$ such that, for $j = 1, \dots, 2^r - 1$, the jth column vector of H is the transpose of the coordinate vector of α^{j-1} relative to the ordered basis $\{1, \alpha, \dots, \alpha^{r-1}\}$ of $\mathbb{F}_{2^r}$ over $\mathbb{F}_2$. Since α is a primitive element of $\mathbb{F}_{2^r}$, the column vectors of H run exactly through all nonzero vectors in $\mathbb{F}_2^r$, and so H is a matrix of the form H_r in Definition 3.5.1 and a parity-check matrix of a code $\mathrm{Ham}(r, 2)$. Furthermore, if $\mathbf{v} = (v_0, v_1, \dots, v_{n-1}) \in \mathbb{F}_2^n$ with $n = 2^r - 1$ and $v(x) = v_0 + v_1 x + \dots + v_{n-1} x^{n-1} \in \mathbb{F}_2[x]$, then $\mathbf{v}H^\top$ is the coordinate vector of $v(\alpha)$ relative to the ordered basis $\{1, \alpha, \dots, \alpha^{r-1}\}$. By Theorem 3.2.37, $\mathbf{v} \in \mathrm{Ham}(r, 2)$ if and only if $\mathbf{v}H^\top = \mathbf{0}$. The latter condition is equivalent to $v(\alpha) = 0$, and this is in turn equivalent to $g(x)$ dividing $v(x)$ in $\mathbb{F}_2[x]$ by Proposition 1.4.38. Our claim is thus established.

Remark 3.5.3 There is a variant of $\mathrm{Ham}(r, 2)$, namely an *extended binary Hamming code* $\overline{\mathrm{Ham}(r, 2)}$. For every integer $r \geq 2$ and each choice of $\mathrm{Ham}(r, 2)$ from the equivalence class, such an extended code is defined by

$$\overline{\mathrm{Ham}(r, 2)} = \left\{ \left(c_1, \dots, c_n, \sum_{j=1}^{n} c_j \right) \in \mathbb{F}_2^{n+1} : (c_1, \dots, c_n) \in \mathrm{Ham}(r, 2) \right\}$$

with $n = 2^r - 1$. The length of $\overline{\mathrm{Ham}(r, 2)}$ is $n + 1 = 2^r$. It is easily seen that $\overline{\mathrm{Ham}(r, 2)}$ is again a linear code. Since the codes $\overline{\mathrm{Ham}(r, 2)}$ and $\mathrm{Ham}(r, 2)$ have the same number of codewords, their dimensions agree, and so $\overline{\mathrm{Ham}(r, 2)}$ has dimension $2^r - 1 - r$. If $(c_1, \dots, c_n) \in \mathrm{Ham}(r, 2)$ has the minimum nonzero Hamming weight 3, then $\left(c_1, \dots, c_n, \sum_{j=1}^n c_j\right)$ has Hamming weight 4, and so $\overline{\mathrm{Ham}(r, 2)}$ has minimum distance 4. In summary, $\overline{\mathrm{Ham}(r, 2)}$ is a binary linear $[2^r, 2^r - 1 - r, 4]$ code.

Example 3.5.4 Consider $\mathrm{Ham}(r, 2)$ with $r = 3$ which is a binary linear $[7, 4, 3]$ code. As we have noted, this code was already discussed in Example 3.2.53. From this example we get a generator matrix

$$G = \begin{pmatrix} 1\,0\,0\,0\,0\,1\,1 \\ 0\,1\,0\,0\,1\,0\,1 \\ 0\,0\,1\,0\,1\,1\,1 \\ 0\,0\,0\,1\,1\,1\,0 \end{pmatrix}$$

of $\mathrm{Ham}(3,2)$. By construction, a generator matrix $\overline{G}$ of $C = \overline{\mathrm{Ham}(3,2)}$ is then given by

$$\overline{G} = \begin{pmatrix} 1\,0\,0\,0\,0\,1\,1\,1 \\ 0\,1\,0\,0\,1\,0\,1\,1 \\ 0\,0\,1\,0\,1\,1\,1\,0 \\ 0\,0\,0\,1\,1\,1\,0\,1 \end{pmatrix}.$$

By Remark 3.5.3, C is a binary linear $[8,4,4]$ code. If $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \mathbf{b}_4 \in \mathbb{F}_2^8$ are the four row vectors of $\overline{G}$, then it is easily verified that $\mathbf{b}_i \cdot \mathbf{b}_j = 0$ for all $i, j = 1, 2, 3, 4$, and so the bilinearity of the dot product implies that $\mathbf{u} \cdot \mathbf{v} = 0$ for all $\mathbf{u}, \mathbf{v} \in C$. This shows that $C \subseteq C^\perp$. Furthermore, $\dim(C) = 4 = 8 - 4 = \dim(C^\perp)$ by Theorem 3.2.34, and so $C = C^\perp$. In other words, $C = \overline{\mathrm{Ham}(3,2)}$ is a self-dual code (see Definition 3.2.57).

How can we generalize Hamming codes from the binary case to the q-ary case for any prime power q? Obviously, we should set up a suitable parity-check matrix over $\mathbb{F}_q$. In a simple-minded generalization of the construction in Definition 3.5.1, we would list all nonzero vectors from $\mathbb{F}_q^r$ as columns. For instance, if $q = 3$ and $r = 2$, this would yield the parity-check matrix

$$H = \begin{pmatrix} 0\,0\,1\,1\,1\,2\,2\,2 \\ 1\,2\,0\,1\,2\,0\,1\,2 \end{pmatrix}.$$

But since the second column vector of H is a scalar multiple of the first column vector of H, the resulting linear code has minimum distance 2 by Corollary 3.2.45. However, we prefer a code with minimum distance 3 like a binary Hamming code since this guarantees that the code is 1-error-correcting. The way to achieve this is to avoid scalar multiples of already chosen column vectors in the parity-check matrix.

Having clarified our goal, we now proceed as follows. For an integer $r \geq 2$ and a finite field $\mathbb{F}_q$, we consider two nonzero vectors from $\mathbb{F}_q^r$ equivalent if one is a scalar multiple of the other. This yields exactly $(q^r - 1)/(q - 1)$ corresponding equivalence classes. We set up an $r \times [(q^r - 1)/(q - 1)]$ matrix $H_{r,q}$ over $\mathbb{F}_q$ by choosing as its column vectors one vector from each of the $(q^r - 1)/(q - 1)$ equivalence classes. Among the column vectors of $H_{r,q}$ we find in particular r nonequivalent vectors of Hamming weight 1, and so the row vectors of $H_{r,q}$ are linearly independent over $\mathbb{F}_q$. Therefore $H_{r,q}$ can be taken as a parity-check matrix of a linear code over $\mathbb{F}_q$. The length of the resulting code is $n = (q^r - 1)/(q - 1)$ and its dimension is $n - r = (q^r - 1)/(q - 1) - r$. This construction can be expressed equivalently in the following form.

Definition 3.5.5 For an integer $r \geq 2$ and a finite field $\mathbb{F}_q$, let $H_{r,q}$ be an $r \times [(q^r - 1)/(q - 1)]$ matrix over $\mathbb{F}_q$ that is obtained by choosing as its column vectors one nonzero vector from each of the $(q^r-1)/(q-1)$ different one-dimensional subspaces of $\mathbb{F}_q^r$. Then the linear $[(q^r-1)/(q-1), (q^r-1)/(q-1)-r]$ code over $\mathbb{F}_q$ with parity-check matrix $H_{r,q}$ is called a *Hamming code* $\mathrm{Ham}(r,q)$ over $\mathbb{F}_q$.

The order of the columns of $H_{r,q}$ and the specific choices of vectors from the one-dimensional subspaces of $\mathbb{F}_q^r$ have not been fixed, and so $\mathrm{Ham}(r,q)$ actually represents a family of codes with the same basic properties. A practical way to write down $H_{r,q}$ is to choose as its column vectors all nonzero vectors from $\mathbb{F}_q^r$ whose first nonzero entry is 1.

Example 3.5.6 Take $r=2$ and $q=3$. We list all nonzero vectors from $\mathbb{F}_3^2$ whose first nonzero entry is 1 and obtain

$$H_{2,3} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 2 & 1 & 0 & 1 \end{pmatrix}.$$

The code $\mathrm{Ham}(2,3)$ with parity-check matrix $H_{2,3}$ is a linear $[4,2]$ code over $\mathbb{F}_3$. Since $H_{2,3}$ is in standard form, a generator matrix of $\mathrm{Ham}(2,3)$ is given by

$$G = \begin{pmatrix} 1 & 0 & 2 & 1 \\ 0 & 1 & 2 & 2 \end{pmatrix}.$$

One shows as in Example 3.5.4 that $\mathrm{Ham}(2,3)$ is a self-dual code.

Theorem 3.5.7 *For every integer $r \geq 2$ and every finite field $\mathbb{F}_q$, any Hamming code* $\mathrm{Ham}(r,q)$ *has minimum distance* 3 *and is perfect.*

Proof. By construction, $C = \mathrm{Ham}(r,q)$ satisfies $d(C) \geq 3$ on account of Theorem 3.2.44. Among the column vectors of $H_{r,q}$ we find $\mathbf{h}_1^\top$, $\mathbf{h}_2^\top$, and $\mathbf{h}_3^\top$, where $\mathbf{h}_1 = (a,0,0,\dots,0) \in \mathbb{F}_q^r$, $\mathbf{h}_2 = (0,b,0,\dots,0) \in \mathbb{F}_q^r$, and $\mathbf{h}_3 = (c,c,0,\dots,0) \in \mathbb{F}_q^r$ for some $a,b,c \in \mathbb{F}_q^*$. Then $a^{-1}\mathbf{h}_1 + b^{-1}\mathbf{h}_2 - c^{-1}\mathbf{h}_3 = \mathbf{0}$, and so $\mathbf{h}_1$, $\mathbf{h}_2$, and $\mathbf{h}_3$ are linearly dependent over $\mathbb{F}_q$. Therefore $d(C)=3$ by Corollary 3.2.45.

According to Definition 3.4.8, C is perfect if

$$\sum_{i=0}^{1} \binom{n}{i} (q-1)^i = q^{n-k}, \tag{3.17}$$

where $n = (q^r-1)/(q-1)$ and $k = (q^r-1)/(q-1)-r$. The left-hand side of (3.17) is equal to $1+n(q-1) = q^r$, and this agrees with the right-hand side of (3.17). □

The parameters of $\mathrm{Ham}(r,q)$ in Definition 3.5.5 indicate that $\mathrm{Ham}(r,q)$ has exactly q^r cosets. Note that in $\mathbb{F}_q^n$ with $n = (q^r-1)/(q-1)$, there are exactly $1+n(q-1)=q^r$ words of Hamming weight at most 1. Different words of Hamming weight at most 1 are in different cosets of $\mathrm{Ham}(r,q)$ since $\mathrm{Ham}(r,q)$ has minimum distance 3 by Theorem 3.5.7. Therefore the unique coset leaders of $\text{Ham}(r,q)$ are exactly the words $\mathbf{0} \in \mathbb{F}_q^n$ and $\mathbf{e}_{j,b} \in \mathbb{F}_q^n$ for $1 \leq j \leq n$ and $b \in \mathbb{F}_q^*$, where $\mathbf{e}_{j,b}$ is the word whose jth coordinate is b and all other coordinates are 0.

The syndrome decoding algorithm (see Algorithm 3.2.52) now attains a particularly simple form for Hamming codes. In view of the preceding discussion, all possible syndromes of $\text{Ham}(r,q)$ are given by $\mathbf{0} \in \mathbb{F}_q^r$ and $S(\mathbf{e}_{j,b}) = b\mathbf{h}_j^\top \in \mathbb{F}_q^r$ for $1 \leq j \leq n$ and $b \in \mathbb{F}_q^*$, where $\mathbf{h}_j$ denotes the jth column vector of $H_{r,q}$.

Algorithm 3.5.8 (Syndrome Decoding Algorithm for Hamming Codes) Let a Hamming code $\text{Ham}(r,q)$ over $\mathbb{F}_q$ with parity-check matrix $H_{r,q}$ and length $n = (q^r-1)/(q-1)$ be given.
Step 1: for a received word $\mathbf{v} \in \mathbb{F}_q^n$, compute the syndrome $S(\mathbf{v}) = \mathbf{v}H_{r,q}^\top$.
Step 2: if $S(\mathbf{v}) = \mathbf{0}$, then assume that no errors have occurred and $\mathbf{v}$ is the most likely sent codeword.
Step 3: if $S(\mathbf{v}) \neq \mathbf{0}$, then find the unique column vector $\mathbf{h}_j$ of $H_{r,q}$ such that $S(\mathbf{v})$ is a scalar multiple of $\mathbf{h}_j^\top$, say $S(\mathbf{v}) = b\mathbf{h}_j^\top$ with $b \in \mathbb{F}_q^*$.
Step 4: with j and b from Step 3, $\mathbf{e}_{j,b}$ is the most likely error word and $\mathbf{c}' = \mathbf{v} - \mathbf{e}_{j,b}$ is the most likely sent codeword.

Example 3.5.9 Consider the Hamming code $\text{Ham}(2,3)$ over $\mathbb{F}_3$ in Example 3.5.6. Suppose that the received word is $\mathbf{v} = (0\ 1\ 1\ 0) \in \mathbb{F}_3^4$. Then $S(\mathbf{v}) = \mathbf{v}H_{2,3}^\top = (2\ 1) \in \mathbb{F}_3^2$, and so $S(\mathbf{v})$ is a scalar multiple of the transposed first column vector $\mathbf{h}_1^\top$ of $H_{2,3}$, namely $S(\mathbf{v}) = 2\mathbf{h}_1^\top$. It follows that $\mathbf{e}_{1,2} = (2\ 0\ 0\ 0) \in \mathbb{F}_3^4$ is the most likely error word and

$$\mathbf{c}' = \mathbf{v} - \mathbf{e}_{1,2} = (0\ 1\ 1\ 0) - (2\ 0\ 0\ 0) = (1\ 1\ 1\ 0)$$

is the most likely sent codeword.

Remark 3.5.10 This syndrome decoding algorithm is even simpler for binary Hamming codes $\text{Ham}(r,2)$, since then in Step 3 of Algorithm 3.5.8 we must have $b = 1 \in \mathbb{F}_2$, and so $S(\mathbf{v}) = \mathbf{h}_j^\top$ for a uniquely determined integer j with $1 \leq j \leq 2^r - 1$. If the columns of the parity-check matrix H_r are arranged in lexicographic order as in the matrix H_3 at the beginning of this subsection, then $S(\mathbf{v})$ corresponds to the binary representation of the integer j. The most likely error word is then the word $\mathbf{e}_j$ whose jth coordinate is 1 and all other coordinates are 0.

Example 3.5.11 Consider the binary Hamming code $\text{Ham}(3,2)$ with parity-check matrix

$$H_3 = \begin{pmatrix} 0\,0\,0\,1\,1\,1\,1 \\ 0\,1\,1\,0\,0\,1\,1 \\ 1\,0\,1\,0\,1\,0\,1 \end{pmatrix}.$$

Note that the first column vector $(0\ 0\ 1)^\top$ of H_3 corresponds to the binary representation of $1 = 0 \cdot 2^2 + 0 \cdot 2 + 1 \cdot 1$, the second column vector $(0\ 1\ 0)^\top$ of H_3 corresponds to the binary representation of $2 = 0 \cdot 2^2 + 1 \cdot 2 + 0 \cdot 1$, and so on. Suppose that the received word is $\mathbf{v} = (0\ 1\ 1\ 1\ 0\ 0\ 0) \in \mathbb{F}_2^7$. Then $S(\mathbf{v}) = \mathbf{v}H_3^\top = (1\ 0\ 1) \in \mathbb{F}_2^3$, which corresponds to the binary representation of $j = 5 = 1 \cdot 2^2 + 0 \cdot 2 + 1 \cdot 1$. Therefore the most likely error word is $\mathbf{e}_5 = (0\ 0\ 0\ 0\ 1\ 0\ 0) \in \mathbb{F}_2^7$ and

$$\mathbf{c}' = \mathbf{v} - \mathbf{e}_5 = (0\ 1\ 1\ 1\ 0\ 0\ 0) - (0\ 0\ 0\ 0\ 1\ 0\ 0) = (0\ 1\ 1\ 1\ 1\ 0\ 0)$$

is the most likely sent codeword.

The dual codes of the Hamming codes also have interesting properties. These codes run under a special name as well.

Definition 3.5.12 The dual code of a Hamming code $\mathrm{Ham}(r, q)$ over $\mathbb{F}_q$ is called a *simplex code* $S(r, q)$ over $\mathbb{F}_q$.

Proposition 3.5.13 *For every integer $r \geq 2$ and every finite field $\mathbb{F}_q$, any simplex code $S(r, q)$ is a linear $[(q^r - 1)/(q - 1), r]$ code over $\mathbb{F}_q$.*

Proof. This follows by using the values of the length and of the dimension of $\mathrm{Ham}(r, q)$ in Definition 3.5.5 and then applying Theorem 3.2.34. □

The name "simplex code" stems from the remarkable property of the codes $S(r, q)$ shown in the following theorem, which is reminiscent of the geometric characteristic of a regular simplex in Euclidean space.

Theorem 3.5.14 *For every integer $r \geq 2$ and every prime power q, any simplex code $S(r, q)$ is an equidistant code. In fact, every nonzero codeword in $S(r, q)$ has Hamming weight q^{r-1}. In particular, $S(r, q)$ has minimum distance q^{r-1}.*

Proof. By definition, an $r \times n$ matrix $H_{r,q}$ of the form in Definition 3.5.5 is a generator matrix of $S(r, q)$, where $n = (q^r - 1)/(q - 1)$. Let $\mathbf{h}_1^\top, \ldots, \mathbf{h}_n^\top$ be the column vectors of $H_{r,q}$. Now we fix a nonzero codeword $\mathbf{c} = (c_1, \ldots, c_n) \in S(r, q)$. Then $\mathbf{c} = \mathbf{a}H_{r,q}$ for some $\mathbf{a} \in \mathbb{F}_q^r$ with $\mathbf{a} \neq \mathbf{0}$. It follows that $c_j = \mathbf{a} \cdot \mathbf{h}_j$ for $1 \leq j \leq n$. Thus, the number of j with $1 \leq j \leq n$ and $c_j = 0$ is equal to the number of $\mathbf{h}_j \in U := \{\mathbf{u} \in \mathbb{F}_q^r : \mathbf{a} \cdot \mathbf{u} = 0\}$. From $\mathbf{a} \neq \mathbf{0}$ we deduce that U is an $(r - 1)$-dimensional subspace of $\mathbb{F}_q^r$, and so U contains exactly $(q^{r-1} - 1)/(q - 1)$ one-dimensional subspaces. Therefore

$$w(\mathbf{c}) = n - \frac{q^{r-1} - 1}{q - 1} = \frac{q^r - 1}{q - 1} - \frac{q^{r-1} - 1}{q - 1} = q^{r-1}.$$

The last part of the theorem follows from Theorem 3.2.14. □

Example 3.5.15 Consider the Hamming code $\mathrm{Ham}(2,3)$ in Example 3.5.6. Its dual code $S(2,3)$ has a generator matrix

$$H_{2,3} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 2 & 1 & 0 & 1 \end{pmatrix}.$$

By forming all linear combinations over $\mathbb{F}_3$ of the row vectors of $H_{2,3}$, we obtain all codewords in $S(2,3)$. In this way it can be verified directly that all nonzero codewords in $S(2,3)$ have Hamming weight 3. Since $\mathrm{Ham}(2,3)$ is self-dual, we have $S(2,3) = \mathrm{Ham}(2,3)$ in this case.

Remark 3.5.16 In the binary case $q = 2$, we have demonstrated in Remark 3.5.2 that, for every integer $r \geq 2$, a suitable code $\mathrm{Ham}(r,2)$ is cyclic with generator polynomial $g(x) \in \mathbb{F}_2[x]$, where $g(x)$ is the minimal polynomial of a primitive element of $\mathbb{F}_{2^r}$ over $\mathbb{F}_2$. Its dual code $S(r,2)$ is therefore also cyclic by Proposition 3.3.23, and the parity-check polynomial of $S(r,2)$ is $g(x)$ according to Remark 3.3.27. In other words, $S(r,2)$ is an irreducible cyclic code over $\mathbb{F}_2$ of length $2^r - 1$ (compare with Subsection 3.3.6). It was already proved in Example 3.4.21 with $q = 2$ and by a different method that $S(r,2)$ is an equidistant code.

Remark 3.5.17 According to Definition 3.2.54 and Theorem 3.5.14, the weight enumerator $A(x)$ of any simplex code $S(r,q)$ is given by

$$A(x) = 1 + (q^r - 1)x^{q^{r-1}} \in \mathbb{Z}[x].$$

Since the dual code of $S(r,q)$ is a Hamming code $\mathrm{Ham}(r,q)$, the MacWilliams identity in Theorem 3.2.55 yields the formula

$$\begin{aligned} A^{\perp}(x) &= q^{-r}(1+(q-1)x)^n \left[1 + (q^r - 1)\Big(\frac{1-x}{1+(q-1)x}\Big)^{q^{r-1}}\right] \\ &= q^{-r}\left[(1+(q-1)x)^n + (q^r-1)(1-x)^{q^{r-1}}(1+(q-1)x)^{n-q^{r-1}}\right] \end{aligned}$$

for the weight enumerator $A^{\perp}(x)$ of $\mathrm{Ham}(r,q)$, where $n = (q^r - 1)/(q-1)$. This formula can be used to determine the number of codewords in $\mathrm{Ham}(r,q)$ of a given Hamming weight.

Binary and ternary simplex codes yield interesting examples of self-orthogonal codes. We recall from Definition 3.2.57 that a nontrivial linear code C is self-orthogonal if $C \subseteq C^{\perp}$.

Theorem 3.5.18 *Every binary simplex code $S(r,2)$ with $r \geq 3$ is self-orthogonal.*

Proof. Since $S(r,2)$ is the dual code of $\mathrm{Ham}(r,2)$, we have to prove that $\mathrm{Ham}(r,2)^{\perp} \subseteq \mathrm{Ham}(r,2)$ for $r \geq 3$. By Remark 3.5.2, we can assume that $\mathrm{Ham}(r,2)$ is cyclic with generator polynomial $g(x) \in \mathbb{F}_2[x]$, where $g(x)$ is the minimal polynomial of a primitive element $\alpha \in \mathbb{F}_{2^r}$ over $\mathbb{F}_2$. Theorem 3.3.26 implies that $\mathrm{Ham}(r,2)^\perp$ is cyclic with generator polynomial $h^*(x) \in \mathbb{F}_2[x]$, where $h^*(x)$ is the reciprocal polynomial of $h(x) = (x^n - 1)/g(x)$ with $n = 2^r - 1$. Now $g^*(x)$ is irreducible over $\mathbb{F}_2$ and $g(\alpha) = 0$ implies $g^*(\alpha^{-1}) = 0$. If α^{-1} were a root of $g(x)$, then $\alpha^{-1} = \alpha^{2^j}$ for some $j = 0, 1, \ldots, r-1$ by Proposition 1.4.47, and so $\alpha^{2^j+1} = 1$. It follows that $2^r - 1$ divides $2^j + 1$. But $2^j + 1 \le 2^{r-1} + 1 < 2^r - 1$ for $r \ge 3$, a contradiction. Thus $g(\alpha^{-1}) \ne 0$, and so the two irreducible polynomials $g(x)$ and $g^*(x)$ over $\mathbb{F}_2$ are coprime. Now $g(x)$ divides $x^n - 1 = (1 - x^n)^* = (x^n - 1)^* = g^*(x)h^*(x)$ in $\mathbb{F}_2[x]$, and so $g(x)$ divides $h^*(x)$ in $\mathbb{F}_2[x]$ by Proposition 1.4.17(ii). This shows that $\mathrm{Ham}(r,2)^\perp \subseteq \mathrm{Ham}(r,2)$. □

Note that Theorem 3.5.18 cannot hold for $r = 2$ since $S(2,2)$ has dimension 2, whereas $S(2,2)^\perp$ has dimension 1. We now turn to ternary simplex codes.

Theorem 3.5.19 *Every ternary simplex code $S(r,3)$ with $r \ge 2$ is self-orthogonal.*

Proof. Since $S(r,3)$ is the dual code of $\mathrm{Ham}(r,3)$, the code $S(r,3)$ has a generator matrix of the form $H_{r,3}$ in Definition 3.5.5. Let its row vectors be $\mathbf{b}_1^{(r)}, \ldots, \mathbf{b}_r^{(r)}$. We note that $w(\mathbf{b}_i^{(r)}) = 3^{r-1}$ for $1 \le i \le r$ by Theorem 3.5.14. Together with $a^2 = 1$ for all $a \in \mathbb{F}_3^*$, this implies that $\mathbf{b}_i^{(r)} \cdot \mathbf{b}_i^{(r)} = 3^{r-1} = 0 \in \mathbb{F}_3$ for $1 \le i \le r$. In order to prove that $S(r,3)$ is self-orthogonal, it remains to show that $\mathbf{b}_i^{(r)} \cdot \mathbf{b}_j^{(r)} = 0$ for $1 \le i < j \le r$. For this purpose, we can assume that in each column of $H_{r,3}$ the first nonzero entry is 1, again because of $a^2 = 1$ for all $a \in \mathbb{F}_3^*$. Permuting columns of $H_{r,3}$ will not change the dot products $\mathbf{b}_i^{(r)} \cdot \mathbf{b}_j^{(r)}$, and so we can write the columns of $H_{r,3}$ in any order. Now we proceed by induction on r. For $r = 2$ we can take

$$H_{2,3} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 2 & 1 \end{pmatrix},$$

and then it is clear that $\mathbf{b}_1^{(2)} \cdot \mathbf{b}_2^{(2)} = 0$. Suppose that the property $\mathbf{b}_i^{(r)} \cdot \mathbf{b}_j^{(r)} = 0$ for $1 \le i < j \le r$ has been shown for some $r \ge 2$ and consider the matrix $H_{r+1,3}$. There are exactly 3^r column vectors of $H_{r+1,3}$ of the form $\binom{1}{\mathbf{u}^\top}$, where $\mathbf{u}^\top$ is an arbitrary column vector over $\mathbb{F}_3$ of length r, and the remaining $\frac{1}{2}(3^r - 1)$ column vectors of $H_{r+1,3}$ are of the form $\binom{0}{\mathbf{v}^\top}$, where $\mathbf{v}^\top$ is a column vector of $H_{r,3}$. For the sake of concreteness, we choose the first row vector of $H_{r+1,3}$ as

$$\mathbf{b}_1^{(r+1)} = (1 \ \ldots \ 1 \ 0 \ \ldots \ 0),$$

where the first 3^r coordinates are 1 and the remaining $\frac{1}{2}(3^r - 1)$ coordinates are 0. For $2 \le j \le r+1$, the dot product $\mathbf{b}_1^{(r+1)} \cdot \mathbf{b}_j^{(r+1)}$ is equal to the sum of the $(j-1)$st coordinates of all $\mathbf{u} \in \mathbb{F}_3^r$, which is $3^{r-1}(0+1+2) = 0 \in \mathbb{F}_3$. For $2 \le i < j \le r+1$, we write $\mathbf{b}_i^{(r+1)} = (\mathbf{c}_i^{(r+1)}, \mathbf{d}_i^{(r+1)})$ and $\mathbf{b}_j^{(r+1)} = (\mathbf{c}_j^{(r+1)}, \mathbf{d}_j^{(r+1)})$, where $\mathbf{c}_i^{(r+1)}$, respectively $\mathbf{c}_j^{(r+1)}$, is formed by the first 3^r coordinates of $\mathbf{b}_i^{(r+1)}$, respectively $\mathbf{b}_j^{(r+1)}$. Then

$$\mathbf{b}_i^{(r+1)} \cdot \mathbf{b}_j^{(r+1)} = \mathbf{c}_i^{(r+1)} \cdot \mathbf{c}_j^{(r+1)} + \mathbf{d}_i^{(r+1)} \cdot \mathbf{d}_j^{(r+1)} = \mathbf{c}_i^{(r+1)} \cdot \mathbf{c}_j^{(r+1)}$$

by the induction hypothesis since $\mathbf{d}_i^{(r+1)}$ and $\mathbf{d}_j^{(r+1)}$ are row vectors of $H_{r,3}$. Furthermore,

$$\mathbf{c}_i^{(r+1)} \cdot \mathbf{c}_j^{(r+1)} = 3^{r-2} \sum_{a,b \in \mathbb{F}_3} ab = 3^{r-2} \Big(\sum_{a \in \mathbb{F}_3} a \Big)^2 = 0 \in \mathbb{F}_3.$$

Therefore $\mathbf{b}_i^{(r+1)} \cdot \mathbf{b}_j^{(r+1)} = 0$ and the induction is complete. □

3.5.2 Golay codes

Golay codes are very pretty flowers in the garden of coding theory. There are, up to equivalence, only four (extended) Golay codes, namely the binary Golay code G_{23}, the extended binary Golay code G_{24}, the ternary Golay code G_{11}, and the extended ternary Golay code G_{12}. The subscripts indicate the lengths of these codes. Golay codes were introduced in the brilliant one-page paper [57], but see also Example 6.3.27. The Golay codes G_{23} and G_{11} belong to the exclusive club of perfect codes.

In order to define G_{23}, we start from the factorization

$$x^{23}-1 = (x+1)(x^{11}+x^9+x^7+x^6+x^5+x+1)(x^{11}+x^{10}+x^6+x^5+x^4+x^2+1) \quad (3.18)$$

in $\mathbb{F}_2[x]$ into irreducible polynomials over $\mathbb{F}_2$ (recall that $-1 = 1$ in $\mathbb{F}_2$). The irreducibility over $\mathbb{F}_2$ of the two factors of degree 11 in (3.18) can be verified by consulting published tables of irreducible polynomials (see for example [103, Chapter 10, Table C]).

Definition 3.5.20 The *binary Golay code* G_{23} is the cyclic code over $\mathbb{F}_2$ of length 23 with generator polynomial $g_{23}(x) = x^{11} + x^9 + x^7 + x^6 + x^5 + x + 1 \in \mathbb{F}_2[x]$.

We may also take the second factor of degree 11 in (3.18), which is the reciprocal polynomial of $g_{23}(x)$, as the generator polynomial of a binary cyclic code of length 23, but this yields an equivalent code. It follows from Theorem 3.3.17 that G_{23} is a binary cyclic $[23, 12]$ code.

The *extended binary Golay code* G_{24} is obtained from G_{23} in the same way as an extended binary Hamming code is obtained from a binary Hamming code, namely

$$G_{24} = \Big\{ \Big(c_1, \ldots, c_{23}, \sum_{j=1}^{23} c_j \Big) \in \mathbb{F}_2^{24} : (c_1, \ldots, c_{23}) \in G_{23} \Big\}.$$

It is clear that G_{24} is a binary linear $[24, 12]$ code. The code G_{24} was used in the Voyager space probes that were launched towards Jupiter and Saturn in 1977. Remarkably, Voyager 1 became the first human-made object that left the solar system and entered interstellar space.

It requires some work to determine the minimum distances of G_{23} and G_{24}. We first study the extended binary Golay code G_{24} in more detail.

Proposition 3.5.21 *The extended binary Golay code G_{24} is self-dual.*

Proof. The generator polynomial $g_{23}(x)$ of G_{23} yields a generator matrix of G_{23} according to Theorem 3.3.19. For $i = 1, \dots, 12$, let $\mathbf{b}_i \in \mathbb{F}_2^{23}$ be the ith row vector of this generator matrix. Then it is easily verified that $\mathbf{b}_1 \cdot \mathbf{b}_i = 1$ for $1 \le i \le 12$. Since each $\mathbf{b}_i$ is a cyclic shift of $\mathbf{b}_1$, it follows that $\mathbf{b}_i \cdot \mathbf{b}_j = 1$ for $1 \le i \le j \le 12$. By construction, the vectors $\mathbf{b}'_i = (\mathbf{b}_i, 1) \in \mathbb{F}_2^{24}$, $i = 1, \dots, 12$, are the row vectors of a generator matrix of G_{24}. These vectors satisfy $\mathbf{b}'_i \cdot \mathbf{b}'_j = \mathbf{b}_i \cdot \mathbf{b}_j + 1 = 0$ for $1 \le i \le j \le 12$, and hence $\mathbf{c} \cdot \mathbf{d} = 0$ for all $\mathbf{c}, \mathbf{d} \in G_{24}$ by the bilinearity of the dot product. This means that $G_{24} \subseteq G_{24}^{\perp}$. Since $\dim(G_{24}) = \dim(G_{24}^{\perp}) = 12$, we conclude that $G_{24} = G_{24}^{\perp}$. □

Lemma 3.5.22 *For every integer $n \ge 1$ and all $\mathbf{u} = (u_1, \dots, u_n) \in \mathbb{F}_2^n$ and $\mathbf{v} = (v_1, \dots, v_n) \in \mathbb{F}_2^n$, put*

$$\mathbf{u} \star \mathbf{v} = (u_1 v_1, \dots, u_n v_n) \in \mathbb{F}_2^n.$$

Then the Hamming weight $w(\mathbf{u} + \mathbf{v})$ of $\mathbf{u} + \mathbf{v}$ satisfies

$$w(\mathbf{u} + \mathbf{v}) = w(\mathbf{u}) + w(\mathbf{v}) - 2w(\mathbf{u} \star \mathbf{v}).$$

Proof. As for similar results, it suffices to give the proof for $n = 1$ (compare with the proof of Proposition 3.1.6(iv)). For this case, an easy calculation for all four pairs $(u, v) \in \mathbb{F}_2^2$ verifies the desired formula. □

Lemma 3.5.23 *The Hamming weight of every codeword in G_{24} is a multiple of* 4.

Proof. Every codeword in G_{24} is a sum of some of the vectors $\mathbf{b}'_1, \dots, \mathbf{b}'_{12}$ in the proof of Proposition 3.5.21. For a single vector $\mathbf{b}'_i$, it is clear that $w(\mathbf{b}'_i) = 8$ for $1 \le i \le 12$. For a sum $\mathbf{b}'_i + \mathbf{b}'_j$ $(1 \le i \le j \le 12)$ of two vectors, we obtain $w(\mathbf{b}'_i + \mathbf{b}'_j) = w(\mathbf{b}'_i) + w(\mathbf{b}'_j) - 2w(\mathbf{b}'_i \star \mathbf{b}'_j)$ by Lemma 3.5.22, hence

$$w(\mathbf{b}'_i + \mathbf{b}'_j) \equiv 2w(\mathbf{b}'_i \star \mathbf{b}'_j) \pmod 4.$$

Now $\mathbf{b}'_i \cdot \mathbf{b}'_j = 0$ by Proposition 3.5.21, thus $w(\mathbf{b}'_i \star \mathbf{b}'_j)$ is even, and so $w(\mathbf{b}'_i + \mathbf{b}'_j) \equiv 0 \pmod 4$. We then continue by induction to get the result for any number of summands. □

Theorem 3.5.24 *The extended binary Golay code G_{24} is a self-dual $[24, 12, 8]$ code.*

Proof. It remains to show that $d(G_{24}) = 8$. It follows from (3.18) that $g_{23}(x)$ has a root $\alpha \in \mathbb{F}_{2^{11}}$ which is a primitive 23rd root of unity. Since $g_{23}(x)$ is irreducible over $\mathbb{F}_2$, all roots of $g_{23}(x)$ are given by Proposition 1.4.47, that is, the roots of $g_{23}(x)$ are

$$\alpha, \alpha^2, \alpha^4, \alpha^8, \alpha^{16}, \alpha^{32} = \alpha^9, \alpha^{18}, \alpha^{36} = \alpha^{13}, \alpha^{26} = \alpha^3, \alpha^6, \alpha^{12}.$$

Then Theorem 3.3.34 with $C = G_{23}$, $b = 1$, and $d = 5$ yields $d(G_{24}) \geq d(G_{23}) \geq 5$, and so $d(G_{24}) \geq 8$ by Lemma 3.5.23. On the other hand, there are codewords in G_{24} of Hamming weight 8 (see the proof of Lemma 3.5.23), and so $d(G_{24}) = 8$. □

Theorem 3.5.25 *The binary Golay code G_{23} is a perfect cyclic $[23, 12, 7]$ code.*

Proof. In order to determine $d(G_{23})$, we note that Theorem 3.5.24 and the relationship between G_{23} and G_{24} imply that $d(G_{23}) \geq 7$. On the other hand, there are codewords in G_{23} of Hamming weight 7 (for instance $\mathbf{b}_1$ in the proof of Proposition 3.5.21), and so $d(G_{23}) = 7$.

In order to show that G_{23} is perfect, we need to check equality in (3.15) with $q = 2$, $n = 23$, $|C| = |G_{23}| = 2^{12}$, and $d = 7$. Note that

$$\binom{23}{0} + \binom{23}{1} + \binom{23}{2} + \binom{23}{3} = 1 + 23 + 253 + 1771 = 2048 = 2^{11},$$

and this yields the desired result. □

We introduce the ternary Golay code G_{11} by starting from the factorization

$$x^{11} - 1 = (x + 2)(x^5 + 2x^3 + x^2 + 2x + 2)(x^5 + x^4 + 2x^3 + x^2 + 2) \qquad (3.19)$$

in $\mathbb{F}_3[x]$ into irreducible polynomials over $\mathbb{F}_3$ (recall that $-1 = 2$ in $\mathbb{F}_3$). The irreducibility over $\mathbb{F}_3$ of the two factors of degree 5 in (3.19) can be checked in [103, Chapter 10, Table C].

Definition 3.5.26 The *ternary Golay code* G_{11} is the cyclic code over $\mathbb{F}_3$ of length 11 with generator polynomial $g_{11}(x) = x^5 + 2x^3 + x^2 + 2x + 2 \in \mathbb{F}_3[x]$.

An equivalent code is obtained by taking the second factor of degree 5 in (3.19) as the generator polynomial of a ternary cyclic code of length 11. Theorem 3.3.17 shows that G_{11} is a ternary cyclic $[11, 6]$ code. The *extended ternary Golay code* G_{12} is defined by

$$G_{12} = \Big\{ \Big(c_1, \ldots, c_{11}, -\sum_{j=1}^{11} c_j \Big) \in \mathbb{F}_3^{12} : (c_1, \ldots, c_{11}) \in G_{11} \Big\}.$$

Then G_{12} is a ternary linear $[12, 6]$ code.

Theorem 3.5.27 *The extended ternary Golay code G_{12} is a self-dual* $[12, 6, 6]$ *code.*

Proof. In order to show that G_{12} is self-dual, we proceed as in the proof of Proposition 3.5.21. The generator polynomial $g_{11}(x)$ of G_{11} yields a generator matrix of G_{11} according to Theorem 3.3.19. For $i = 1, \dots, 6$, let $\mathbf{b}_i \in \mathbb{F}_3^{11}$ be the ith row vector of this generator matrix. Then $\mathbf{b}_i \cdot \mathbf{b}_j = 2$ for $1 \le i \le j \le 6$. The vectors $\mathbf{b}'_i = (\mathbf{b}_i, 1) \in \mathbb{F}_3^{12}$, $i = 1, \dots, 6$, are the row vectors of a generator matrix of G_{12}. These vectors satisfy $\mathbf{b}'_i \cdot \mathbf{b}'_j = \mathbf{b}_i \cdot \mathbf{b}_j + 1 = 0$ for $1 \le i \le j \le 6$, and we conclude as in the proof of Proposition 3.5.21 that G_{12} is self-dual.

From (3.19) we see that $g_{11}(x)$ has a root $\alpha \in \mathbb{F}_{3^5}$ which is a primitive 11th root of unity. Since $g_{11}(x)$ is irreducible over $\mathbb{F}_3$, all roots of $g_{11}(x)$ are given by Proposition 1.4.47, that is, the roots of $g_{11}(x)$ are

$$\alpha, \alpha^3, \alpha^9, \alpha^{27} = \alpha^5, \alpha^{15} = \alpha^4.$$

Then Theorem 3.3.34 with $C = G_{11}$, $b = 3$, and $d = 4$ yields $d(G_{12}) \ge d(G_{11}) \ge 4$. Since $\mathbf{c} \cdot \mathbf{c} = 0$ for all $\mathbf{c} \in G_{12}$ and $a^2 = 1$ for $a \in \mathbb{F}_3^*$, the Hamming weight of every $\mathbf{c} \in G_{12}$ is divisible by 3, and so $d(G_{12}) \ge 6$. On the other hand, since $w(\mathbf{b}'_1) = 6$ and $\mathbf{b}'_1 \in G_{12}$, we get $d(G_{12}) = 6$. □

Theorem 3.5.28 *The ternary Golay code G_{11} is a perfect cyclic* $[11, 6, 5]$ *code.*

Proof. Theorem 3.5.27 and the relationship between G_{11} and G_{12} imply that $d(G_{11}) \ge 5$. On the other hand, there are codewords in G_{11} of Hamming weight 5 (for instance $\mathbf{b}_1$ in the proof of Theorem 3.5.27), and so $d(G_{11}) = 5$.

In order to show that G_{11} is perfect, we need to check equality in (3.15) with $q = 3$, $n = 11$, $|C| = |G_{11}| = 3^6$, and $d = 5$. Note that

$$\binom{11}{0} + \binom{11}{1} \cdot 2 + \binom{11}{2} \cdot 2^2 = 1 + 22 + 220 = 243 = 3^5,$$

and this yields the desired result. □

3.5.3 Reed-Solomon codes and BCH codes

We consider further interesting families of linear codes. We start with Reed-Solomon codes which are employed, for instance, in CD players. A Reed-Solomon code over $\mathbb{F}_{256}$ is part of the CCSDS (Consultative Committee for Space Data Systems) standard for space communications. Reed-Solomon codes were first constructed in different incarnations by Bush [17] and Reed and Solomon [165].

Definition 3.5.29 Let $q \geq 3$ be a prime power, let $c \in \mathbb{F}_q^*$ be a primitive element of $\mathbb{F}_q$, and let b and d be integers with $b \geq 0$ and $2 \leq d \leq q-1$. Then the cyclic code over $\mathbb{F}_q$ of length $q-1$ with generator polynomial $g(x) = \prod_{j=b}^{b+d-2}(x - c^j) \in \mathbb{F}_q[x]$ is called a *Reed-Solomon code* over $\mathbb{F}_q$ and denoted by $\mathrm{RS}_q(b, c, d)$.

Theorem 3.5.30 *The Reed-Solomon code* $\mathrm{RS}_q(b, c, d)$ *is a cyclic* $[q-1, q-d, d]$ *code over* $\mathbb{F}_q$ *and an MDS code.*

Proof. For $C = \mathrm{RS}_q(b, c, d)$, Theorem 3.3.17 implies that $\dim(C) = q - 1 - \deg(g(x)) = q-1-(d-1) = q-d$. Since a primitive element of $\mathbb{F}_q$ is a primitive nth root of unity with $n = q-1$, we can apply Theorem 3.3.34 and obtain $d(C) \geq d$. On the other hand, the Singleton bound for linear codes (see Corollary 3.4.11) shows that $d(C) \leq q - 1 - (q-d) + 1 = d$, and so $d(C) = d$. Moreover, C is an MDS code according to Definition 3.4.13. □

Example 3.5.31 Put $q = 7$, $b = 1$, and $d = 3$. Then we can take $c = 3 \in \mathbb{F}_7^*$ as a primitive element of $\mathbb{F}_7$. The cyclic $[6, 4, 3]$ code over $\mathbb{F}_7$ with generator polynomial

$$g(x) = (x-3)(x-3^2) = (x-3)(x-2) = x^2 + 2x + 6 \in \mathbb{F}_7[x]$$

is the Reed-Solomon code $\mathrm{RS}_7(1, 3, 3)$.

The Reed-Solomon codes $\mathrm{RS}_q(b, c, d)$ with $b = 1$ can be represented in the following alternative form. We recall that $\mathbb{F}_q[x]_{<n}$ denotes the set of polynomials over $\mathbb{F}_q$ of degree less than n.

Theorem 3.5.32 *Let* $q \geq 3$ *be a prime power, let* $c \in \mathbb{F}_q^*$ *be a primitive element of* $\mathbb{F}_q$*, and let* d *be an integer with* $2 \leq d \leq q-1$*. Then*

$$\mathrm{RS}_q(1, c, d) = \{\big(f(1), f(c), f(c^2), \dots, f(c^{q-2})\big) \in \mathbb{F}_q^{q-1} : f \in \mathbb{F}_q[x]_{<q-d}\}. \tag{3.20}$$

Proof. Let C be the set on the right-hand side of (3.20). It is clear that C is a linear code over $\mathbb{F}_q$ of length $q-1$. Since the linear transformation

$$f \in \mathbb{F}_q[x]_{<q-d} \mapsto \mathbf{u}_f := \big(f(1), f(c), f(c^2), \dots, f(c^{q-2})\big) \in C$$

is bijective, we obtain $\dim(C) = \dim(\mathbb{F}_q[x]_{<q-d}) = q-d$. Furthermore, C is a cyclic code since for every $f \in \mathbb{F}_q[x]_{<q-d}$ the cyclic shift $\mathbf{u}_f^1$ is equal to $\mathbf{u}_h$ with $h(x) = f(c^{-1}x) \in \mathbb{F}_q[x]_{<q-d}$.

Let $v(x) \in \mathbb{F}_q[x]$ be the generator polynomial of C and let $\pi^{-1}(v(x)) = \mathbf{v} = (v_0, v_1, \dots, v_{q-2}) \in \mathbb{F}_q^{q-1}$ be the vector corresponding to $v(x)$ according to (3.9). Then $\mathbf{v} = \mathbf{u}_f$ for some $f \in \mathbb{F}_q[x]_{<q-d}$. Lemma 3.3.33 shows that the Mattson-Solomon polynomial $M_{\mathbf{v}}(x)$ of $\mathbf{v}$ satisfies

$$M_{\mathbf{v}}(c^i) = -v_i = -f(c^i) \qquad \text{for } i = 0, 1, \dots, q-2.$$

Hence the polynomial $M_{\mathbf{v}}(x)+f(x)$ of degree at most $q-2$ has $q-1$ distinct roots, and so $M_{\mathbf{v}}(x) = -f(x)$. Consequently $\deg(M_{\mathbf{v}}(x)) \leq q-d-1$, and Definition 3.3.32 implies that $v(c^j)=0$ for $j=1,\ldots,d-1$. Therefore $\prod_{j=1}^{d-1}(x-c^j)$ divides $v(x)$ in $\mathbb{F}_q[x]$. Now $\deg(v(x)) = q-1-\dim(C) = d-1$ by Theorem 3.3.17, and so $v(x) = \prod_{j=1}^{d-1}(x-c^j)$. This means that $C = \mathrm{RS}_q(1,c,d)$ according to Definition 3.5.29. □

Remark 3.5.33 The fact that $\mathrm{RS}_q(1,c,d)$ has minimum distance d can be deduced also from Theorem 3.5.32. For every nonzero $f \in \mathbb{F}_q[x]_{<q-d}$, the number of its roots is at most $q-d-1$, and so with the notation in the proof of Theorem 3.5.32 we get $w(\mathbf{u}_f) \geq q-1-(q-d-1) = d$. Therefore $d(\mathrm{RS}_q(1,c,d)) \geq d$. On the other hand, the Singleton bound for linear codes shows as in the proof of Theorem 3.5.30 that $d(\mathrm{RS}_q(1,c,d)) \leq d$, and so $d(\mathrm{RS}_q(1,c,d)) = d$.

Remark 3.5.34 The code in (3.20) can be generalized in a straightforward manner. Let q be a prime power, let n be an integer with $2 \leq n \leq q$, and let k be an integer with $1 \leq k \leq n$. Choose distinct elements $c_1,\ldots,c_n \in \mathbb{F}_q$ and arbitrary nonzero elements $a_1,\ldots,a_n \in \mathbb{F}_q$. Then we introduce the linear code

$$C = \{(a_1 f(c_1),\ldots,a_n f(c_n)) \in \mathbb{F}_q^n : f \in \mathbb{F}_q[x]_{<k}\}.$$

A code of this type is called a *generalized Reed-Solomon code*. It is shown as in Remark 3.5.33 and the proof of Theorem 3.5.32 that C is a linear $[n,k,n-k+1]$ code over $\mathbb{F}_q$. Consequently, every generalized Reed-Solomon code is an MDS code.

Remark 3.5.35 For a Reed-Solomon code $\mathrm{RS}_q(1,c,d)$ as in Theorem 3.5.32, the *extended Reed-Solomon code* is defined by

$$\overline{\mathrm{RS}_q(1,c,d)} = \Big\{\Big(c_0,c_1,\ldots,c_{q-2},-\sum_{j=0}^{q-2} c_j\Big) \in \mathbb{F}_q^q : (c_0,c_1,\ldots,c_{q-2}) \in \mathrm{RS}_q(1,c,d)\Big\}.$$

By Theorem 3.5.30 it is obvious that $\overline{\mathrm{RS}_q(1,c,d)}$ is a linear $[q,q-d]$ code over $\mathbb{F}_q$. Since $\sum_{j=0}^{q-2} f(c^j) = -f(0)$ for all primitive elements c of $\mathbb{F}_q$ and all $f \in \mathbb{F}_q[x]_{<q-d}$, it follows from Theorem 3.5.32 that every codeword in $\overline{\mathrm{RS}_q(1,c,d)}$ has the form

$$\big(f(1),f(c),f(c^2),\ldots,f(c^{q-2}),f(0)\big) \in \mathbb{F}_q^q \qquad \text{for some } f \in \mathbb{F}_q[x]_{<q-d}.$$

It is then proved as in Remark 3.5.33 that $\overline{\mathrm{RS}_q(1,c,d)}$ has minimum distance $d+1$, and so $\overline{\mathrm{RS}_q(1,c,d)}$ is an MDS code. Alternatively, the parameters of $\overline{\mathrm{RS}_q(1,c,d)}$ can be obtained by noting that this code is a generalized Reed-Solomon code (see Remark 3.5.34).

Now we generalize Reed-Solomon codes in a different direction. We noted in Subsection 3.3.5 that cyclic codes over $\mathbb{F}_q$ can be defined by means of roots of polynomials over $\mathbb{F}_q$. The Reed-Solomon code $\mathrm{RS}_q(b, c, d)$ in Definition 3.5.29 is determined via the roots $c^b, c^{b+1}, \ldots, c^{b+d-2}$ of its generator polynomial $g(x)$, where $c \in \mathbb{F}_q^*$ is a primitive element of $\mathbb{F}_q$, that is, a primitive $(q-1)$st root of unity. For an arbitrary finite field $\mathbb{F}_q$ and an integer $n \geq 2$ with $\gcd(n, q) = 1$, let now γ be a primitive nth root of unity in a finite extension field of $\mathbb{F}_q$ (see Subsection 3.3.5). For given integers b and d with $b \geq 0$ and $2 \leq d \leq n$, we consider the cyclic code over $\mathbb{F}_q$ determined by the roots $\gamma^b, \gamma^{b+1}, \ldots, \gamma^{b+d-2}$. More precisely, for $i = b, b+1, \ldots, b+d-2$, let $m_i(x) \in \mathbb{F}_q[x]$ be the minimal polynomial of γ^i over $\mathbb{F}_q$; then the generator polynomial of the cyclic code is obtained from (3.12).

Definition 3.5.36 Let q be an arbitrary prime power and let $n \geq 2$ be an integer with $\gcd(n, q) = 1$. Let b and d be integers with $b \geq 0$ and $2 \leq d \leq n$. Then the cyclic code over $\mathbb{F}_q$ of length n with generator polynomial

$$g(x) = \mathrm{lcm}(m_b(x), m_{b+1}(x), \ldots, m_{b+d-2}(x)) \in \mathbb{F}_q[x], \tag{3.21}$$

where we assume that $\deg(g(x)) < n$, is called a *BCH code* over $\mathbb{F}_q$ of designed distance d. Here $m_i(x) \in \mathbb{F}_q[x]$ is the minimal polynomial of γ^i over $\mathbb{F}_q$ for $b \leq i \leq b+d-2$ and γ is a primitive nth root of unity in a finite extension field of $\mathbb{F}_q$.

The acronym BCH stems from the initials of the inventors of BCH codes, namely Bose and Ray-Chaudhuri [13] and Hocquenghem [65]. BCH codes over $\mathbb{F}_q$ are very popular in practical applications since they permit an efficient decoding algorithm and, for every fixed q, we can achieve an arbitrarily large minimum distance by a suitable choice of the parameters.

Theorem 3.5.37 *Every BCH code C of designed distance d has minimum distance $d(C) \geq d$.*

Proof. This follows immediately from Theorem 3.3.34 and Definition 3.5.36. □

Example 3.5.38 Let $q = 2$, $n = 15$, $b = 1$, and $d = 5$. A suitable primitive 15th root of unity is obtained by choosing a root $\gamma \in \mathbb{F}_{16}$ of the irreducible polynomial $x^4 + x + 1$ over $\mathbb{F}_2$. Then with the notation of Definition 3.5.36, we get $m_1(x) = m_2(x) = m_4(x) = x^4 + x + 1 \in \mathbb{F}_2[x]$ and $m_3(x) = x^4 + x^3 + x^2 + x + 1 \in \mathbb{F}_2[x]$. Thus, the corresponding binary BCH code C is the binary cyclic code of length 15 with generator polynomial

$$g(x) = m_1(x)m_3(x) = x^8 + x^7 + x^6 + x^4 + 1 \in \mathbb{F}_2[x]$$

obtained by (3.21). It follows from Theorem 3.3.17 that $\dim(C) = 7$. Furthermore, Theorem 3.5.37 shows that $d(C) \geq 5$. Since the codeword in C corresponding to $g(x)$ according to (3.9) has Hamming weight 5, it follows that $d(C) = 5$. Therefore C is a 2-error-correcting code by Theorem 3.1.14.

Example 3.5.39 The true minimum distance of a BCH code can be larger than its designed distance. Let $q = 2$, $n = 23$, $b = 1$, and $d = 5$. It was noted in the proof of Theorem 3.5.24 that a primitive 23rd root of unity is obtained as a root of the polynomial $g_{23}(x) \in \mathbb{F}_2[x]$ in Definition 3.5.20. Again by the proof of Theorem 3.5.24, we have $m_1(x) = m_2(x) = m_3(x) = m_4(x) = g_{23}(x)$, and so the corresponding binary BCH code C is the binary cyclic code of length 23 with generator polynomial $g_{23}(x)$ according to (3.21). In other words, C is the binary Golay code G_{23} (see Definition 3.5.20). The designed distance of C is 5, but its true minimum distance is 7 according to Theorem 3.5.25.

Theorem 3.5.40 *The dimension k of a BCH code over $\mathbb{F}_q$ of length n and designed distance d satisfies $k \geq n - (d-1)h$, where h is the multiplicative order of q modulo n. If $q = 2$, $d = 2t + 1$ is odd, and $b = 1$, then $k \geq n - th$.*

Proof. In view of Theorem 3.3.17, it suffices to prove that the polynomial $g(x)$ in (3.21) satisfies $\deg(g(x)) \leq (d-1)h$ in the first case and $\deg(g(x)) \leq th$ in the second case. It was shown in the beginning of Subsection 3.3.6 that there exists a primitive nth root of unity $\gamma \in \mathbb{F}_{q^h}$. Then for every integer $i \geq 0$, the element γ^i is also in $\mathbb{F}_{q^h}$, and so $\deg(m_i(x)) \leq h$. It follows that

$$\deg(g(x)) \leq \sum_{i=b}^{b+d-2} \deg(m_i(x)) \leq (d-1)h.$$

In the second case, with every γ^i the element γ^{2i} is also a root of $m_i(x)$ (see Proposition 1.4.47), and so $m_{2i}(x) = m_i(x)$. Therefore

$$g(x) = \mathrm{lcm}(m_1(x), m_3(x), m_5(x), \ldots, m_{2t-1}(x)),$$

hence

$$\deg(g(x)) \leq \sum_{i=0}^{t-1} \deg(m_{2i+1}(x)) \leq th,$$

and the theorem is proved in all cases. □

Remark 3.5.41 We demonstrated in Remark 3.5.2 that for every integer $r \geq 2$, a suitable binary Hamming code $\mathrm{Ham}(r, 2)$ is cyclic with generator polynomial $g(x) \in \mathbb{F}_2[x]$, where $g(x)$ is the minimal polynomial of a primitive nth root of unity over $\mathbb{F}_2$ with $n = 2^r - 1$. If with $q = 2$ and $n = 2^r - 1$ we put $b = 1$ and $d = 3$ in

Definition 3.5.36, then $m_1(x) = m_2(x) = g(x)$, and so $\mathrm{Ham}(r, 2)$ can also be viewed as a binary BCH code of length $2^r - 1$ and designed distance 3. For $q \geq 3$, some Hamming codes over $\mathbb{F}_q$ can again be interpreted as BCH codes. Let $r \geq 2$ be an integer with $\gcd(r, q-1) = 1$ and put $n = (q^r - 1)/(q-1)$. Then

$$n = \sum_{s=0}^{r-1} q^s \equiv \sum_{s=0}^{r-1} 1 \equiv r \pmod{q-1},$$

and so $\gcd(n, q-1) = 1$. We choose a primitive element β of $\mathbb{F}_{q^r}$, and then $\gamma = \beta^{q-1} \in \mathbb{F}_{q^r}$ is a primitive nth root of unity. We claim that for each choice of integers u and w with $0 \leq u < w \leq n-1$, there is no $c \in \mathbb{F}_q^*$ with $\gamma^w = c\gamma^u$. For otherwise $\gamma^{w-u} = c$, hence $\gamma^{(w-u)(q-1)} = c^{q-1} = 1$, and so n must divide $(w-u)(q-1)$. But $\gcd(n, q-1) = 1$, which implies that n divides $w-u$, a contradiction. Consequently, the $r \times n$ matrix $H_{r,q}$ with columns $1, \gamma, \gamma^2, \ldots, \gamma^{n-1}$ is of the type in Definition 3.5.5, where by a column γ^j (with $0 \leq j \leq n-1$) we mean the transpose of the coordinate vector of γ^j relative to a fixed ordered basis of $\mathbb{F}_{q^r}$ over $\mathbb{F}_q$. The linear code over $\mathbb{F}_q$ with parity-check matrix $H_{r,q}$ is thus a Hamming code $\mathrm{Ham}(r, q)$ over $\mathbb{F}_q$. Now we consider the BCH code C over $\mathbb{F}_q$ of length n with $b = 1$, $d = 2$, and primitive nth root of unity γ. Then C is cyclic with generator polynomial $g(x) \in \mathbb{F}_q[x]$, where $g(x)$ is the minimal polynomial of γ over $\mathbb{F}_q$. For $\mathbf{v} = (v_0, v_1, \ldots, v_{n-1}) \in \mathbb{F}_q^n$, the identity $\mathbf{v}H_{r,q}^\top = \mathbf{0}$ holds if and only if the corresponding polynomial $v(x) = v_0 + v_1 x + \cdots + v_{n-1}x^{n-1} \in \mathbb{F}_q[x]$ satisfies $v(\gamma) = 0$, and this is in turn equivalent to $g(x)$ dividing $v(x)$ in $\mathbb{F}_q[x]$ by Proposition 1.4.38. Therefore $\mathrm{Ham}(r, q)$ is equal to the BCH code C. Note that for $q = 2$ the condition $\gcd(r, q-1) = 1$ is satisfied for all integers $r \geq 2$, and so for every $r \geq 2$ a suitable code $\mathrm{Ham}(r, 2)$ can always be interpreted as a BCH code.

3.6 A glimpse of advanced topics

There are various generalizations of the concept of a cyclic code. A linear code $C \subseteq \mathbb{F}_q^n$ is *constacyclic* if there exists a constant element $a \in \mathbb{F}_q^*$ such that, for every $(c_0, c_1, \ldots, c_{n-1}) \in C$, the word $(ac_{n-1}, c_0, \ldots, c_{n-2})$ is also in C. Cyclic codes correspond to the special case $a = 1 \in \mathbb{F}_q^*$. The analog of Theorem 3.3.11 says that a subset C of $\mathbb{F}_q^n$ is a constacyclic code (for the element $a \in \mathbb{F}_q^*$) if and only if $\pi(C)$ is a nonzero ideal of the residue class ring $\mathbb{F}_q[x]/(x^n - a)$. A discussion of constacyclic codes can be found in Aydin and Asamov [5]. For a length $n \geq 2$ and a proper divisor l of n, a linear code $C \subseteq \mathbb{F}_q^n$ is *quasicyclic* (of index l) if the cyclic shift $\mathbf{c}^l$ is in C for every $\mathbf{c} \in C$. The case $l = 1$ yields cyclic codes. The family of quasicyclic codes is much wider than that of cyclic codes and contains many good codes. A structure theory of quasicyclic codes was developed by Ling and Solé [105] (see also [93] and [104]). An expository account of quasicyclic codes is given in the recent book of Baldi [8, Chapter 3].

A substantial part of coding theory is devoted to finding bounds for the parameters of codes, especially for linear codes. Besides the bounds in Section 3.4, another classical bound is the *Griesmer bound* which says that every linear $[n, k, d]$ code over $\mathbb{F}_q$ satisfies

$$n \geq \sum_{i=0}^{k-1} \lceil d/q^i \rceil .$$

A proof of this bound is given in [106, Section 5.7]. Since $\lceil d/q^0 \rceil = d$ and $\lceil d/q^i \rceil \geq 1$ for $i = 1, \ldots, k-1$, the Griesmer bound implies the Singleton bound for linear codes (see Corollary 3.4.11 for the latter bound). The linear $[(q^r - 1)/(q-1), r, q^{r-1}]$ simplex code $S(r, q)$ over $\mathbb{F}_q$ (see Definition 3.5.12) and Reed-Solomon codes over $\mathbb{F}_q$ show that we can have equality in the Griesmer bound. Surveys of various other bounds, such as the so-called linear programming bounds based on optimization methods, are presented in [108, Chapter 17] and [162, Chapter 4].

Another direction in which bounds in coding theory have been explored is the asymptotic theory of codes. Here one studies the behavior of code parameters as the length of the underlying codes tends to infinity. One may consider arbitrary codes in this theory, but we focus on the case of linear codes. It is customary in this theory to relate the dimension $k(C)$ and the minimum distance $d(C)$ of a linear code C to the length $n(C)$ of C, and so we speak of the *information rate* $k(C)/n(C)$ and the *relative minimum distance* $d(C)/n(C)$ of C. Obviously, the information rate and the relative minimum distance belong to the unit interval $[0, 1]$. The basic object in the asymptotic theory of linear codes is the following set of ordered pairs of asymptotic relative minimum distances and asymptotic information rates. For a fixed prime power q, let $\mathcal{U}_q$ be the set of points (δ, R) in the unit square $[0, 1]^2$ for which there exists a sequence $C_1, C_2, \ldots$ of linear codes over $\mathbb{F}_q$ with $n(C_i) \to \infty$ as $i \to \infty$ and

$$\lim_{i\to\infty} \frac{d(C_i)}{n(C_i)} = \delta, \quad \lim_{i\to\infty} \frac{k(C_i)}{n(C_i)} = R.$$

Then the function α_q on $[0, 1]$ is defined by

$$\alpha_q(\delta) = \sup \{R \in [0, 1] : (\delta, R) \in \mathcal{U}_q\} \qquad \text{for } 0 \leq \delta \leq 1.$$

Thus, $\alpha_q(\delta)$ is the largest asymptotic information rate that can be achieved for a given asymptotic relative minimum distance δ of linear codes over $\mathbb{F}_q$ of increasing length. It can be shown that $\mathcal{U}_q$ is the set of points in the first quadrant of the Euclidean plane lying under or on the graph of α_q. Consequently, $\mathcal{U}_q$ is completely determined by the function α_q.

The study of the function α_q is a fascinating topic in coding theory. It is known that α_q is a nonincreasing continuous function on $[0, 1]$ with $\alpha_q(0) = 1$. It follows from the Plotkin bound in Theorem 3.4.19 that $\alpha_q(\delta) = 0$ for $(q-1)/q < \delta \leq 1$, and the continuity of α_q yields $\alpha_q((q-1)/q) = 0$. The function α_q is not known explicitly on the open interval $(0,(q-1)/q)$. The next best thing is then to give lower bounds on $\alpha_q(\delta)$ for $0<\delta<(q-1)/q$, so that for a given δ in this range we can identify at least some asymptotic information rates that are attainable. The classical lower bound on α_q is the *asymptotic Gilbert-Varshamov bound*

$$\alpha_q(\delta)\geq 1-\delta\log_q(q-1)+\delta\log_q\delta+(1-\delta)\log_q(1-\delta) \qquad \text{for } 0<\delta<\frac{q-1}{q}, \quad (3.22)$$

where $\log_q$ denotes the logarithm to the base q. A derivation of the asymptotic Gilbert-Varshamov bound from the Gilbert-Varshamov bound in Theorem 3.4.3 is presented for example in [148, Section 5.3]. The only way that is currently known to improve on the bound (3.22) is by means of the sophisticated theory of algebraic-geometry codes (see below). Standard families of elementary linear codes such as BCH codes yield good codes of relatively short lengths, but they tend to be useless in the asymptotic theory (compare with [108, Section 9.5]).

We mentioned in Subsection 3.5.3 that BCH codes permit an efficient decoding algorithm. In fact, the structure of BCH codes yields information about the syndromes of received words. For instance, in a suitable interpretation the coordinates of the syndrome satisfy a linear recurrence relation. The coefficients of this linear recurrence relation allow the determination of the error locations. The computation of the desired linear recurrence relation is accomplished by the Berlekamp-Massey algorithm or by the Euclidean algorithm for polynomials over finite fields. We refer to [103, Sections 8.2 and 8.3], [108, Section 9.6], and [162, Chapter 19] for detailed descriptions of the decoding algorithm for BCH codes.

The binary Golay code G_{23} and the ternary Golay code G_{11} belong to the family of quadratic-residue codes. Consider first the cyclic code G_{23} with generator polynomial $g_{23}(x)\in\mathbb{F}_2[x]$ given in Definition 3.5.20. The roots of $g_{23}(x)$ are powers of the primitive 23rd root of unity $\alpha\in F_{2^{11}}$ and they are listed in the proof of Theorem 3.5.24. The exponents on α that yield roots of $g_{23}(x)$ are, in increasing order, given by $1,2,3,4,6,8,9,12,13,16,18$. The crucial observation is now that these numbers are exactly all quadratic residues modulo 23 in the least residue system modulo 23. Similarly, for G_{11} with generator polynomial $g_{11}(x)\in\mathbb{F}_3[x]$ given in Definition 3.5.26, the exponents on the primitive 11th root of unity $\alpha\in\mathbb{F}_{3^5}$ that yield roots of $g_{11}(x)$ are $1,3,4,5,9$ (see the proof of Theorem 3.5.27). These numbers are exactly all quadratic residues modulo 11 in the least residue system modulo 11. The general definition of a *quadratic-residue code* is now as follows. The length of a quadratic-residue code over $\mathbb{F}_q$ is an odd prime number n with $\gcd(n,q)=1$. Furthermore, it is assumed that q is a quadratic residue modulo n. Let S_n be the set of all quadratic residues modulo n in the least residue system modulo n and let α be a primitive nth root of unity in a finite extension field of $\mathbb{F}_q$. Since q is a quadratic residue modulo n, it is easily seen that the polynomial

$$g(x)=\prod_{s\in S_n}(x-\alpha^s)$$

belongs to $\mathbb{F}_q[x]$. The cyclic code over $\mathbb{F}_q$ of length n and with generator polynomial $g(x)$ is, by definition, a quadratic-residue code. It is a cyclic $[n,(n+1)/2]$ code. The same construction works if S_n is replaced by the set N_n of all quadratic nonresidues modulo n in the least residue system modulo n. Expositions of the theory of quadratic-residue codes can be found in [106, Section 8.3] and [108, Chapter 16].

Another important family of special linear codes is that of Reed-Muller codes which is considered mainly in the binary case. Binary Reed-Muller codes are defined for every order $m \geq 1$. A *binary first-order Reed-Muller code* $\mathcal{R}(1,r)$ with an integer $r \geq 2$ is simply the dual code of an extended binary Hamming code $\overline{\mathrm{Ham}(r,2)}$ (see Remark 3.5.3 for the latter codes). For $r=1$ we formally put $\mathcal{R}(1,1)=\mathbb{F}_2^2$. Every code $\mathcal{R}(1,r)$ is a linear $[2^r, r+1]$ code. Generator matrices of (suitable choices of) $\mathcal{R}(1,r)$ can be obtained recursively. For $r=1$ a generator matrix of $\mathcal{R}(1,1)$ is

$$G_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

If the $(r+1)\times 2^r$ matrix G_r over $\mathbb{F}_2$ is a generator matrix of $\mathcal{R}(1,r)$ for some $r \geq 1$, then a generator matrix of $\mathcal{R}(1,r+1)$ is the $(r+2)\times 2^{r+1}$ matrix

$$G_{r+1} = \begin{pmatrix} G_r & G_r \\ \mathbf{0} & \mathbf{1} \end{pmatrix}$$

over $\mathbb{F}_2$, where $\mathbf{0}=(0,\dots,0)\in\mathbb{F}_2^{2^r}$ and $\mathbf{1}=(1,\dots,1)\in\mathbb{F}_2^{2^r}$. For every $r\geq 1$, the minimum distance of $\mathcal{R}(1,r)$ is 2^{r-1}, and in fact all codewords in $\mathcal{R}(1,r)$ except $\mathbf{0}$ and $\mathbf{1}$ have Hamming weight 2^{r-1}. A binary Reed-Muller code $\mathcal{R}(1,5)$ was used in the Mariner space probes that were launched towards Mars in 1969 and 1971. The binary Reed-Muller codes $\mathcal{R}(m,r)$ of order $m\geq 2$ are defined by a double recursion on m and r. For all integers $m\geq 1$ and $r\geq m$, $\mathcal{R}(m,r)$ is a linear $[2^r,k,2^{r-m}]$ code over $\mathbb{F}_2$ with $k=\sum_{j=0}^{m}\binom{r}{j}$. We refer to [108, Chapters 13–15] and [162, Chapter 16] for informative accounts of Reed-Muller codes.

A far-reaching generalization of Reed-Solomon codes leads to the family of algebraic-geometry codes. As our starting point we take the generalized Reed-Solomon code C in Remark 3.5.34 with $a_j=1\in\mathbb{F}_q$ for $1\leq j\leq n$. The jth coordinate of a typical codeword in C is $f(c_j)$ with $c_j\in\mathbb{F}_q$. Note that $f(x)-f(c_j)$ is divisible by $x-c_j$ in $\mathbb{F}_q[x]$, and indeed $f(c_j)$ is the unique element $b\in\mathbb{F}_q$ such that $x-c_j$ divides $f(x)-b$ in $\mathbb{F}_q[x]$, or equivalently $\nu_{x-c_j}(f(x)-b)\geq 1$. Here ν_{x-c_j} is the valuation of the rational function field $\mathbb{F}_q(x)$ (that is, the field of fractions of polynomials over $\mathbb{F}_q$) defined as follows: $\nu_{x-c_j}(0)=\infty$, and for a nonzero $f(x)\in\mathbb{F}_q(x)$ we put $\nu_{x-c_j}(f(x))=e$, where e is the unique integer such that $f(x)=(x-c_j)^e h(x)$ with $h(x)\in\mathbb{F}_q(x)$ and $x-c_j$ dividing neither the numerator nor the denominator of $h(x)$. This definition can be generalized by replacing the linear polynomial $x-c_j$ by any monic irreducible polynomial $p(x)$ over $\mathbb{F}_q$, and this yields the valuation $\nu_{p(x)}$ of $\mathbb{F}_q(x)$. There are two conditions in the definition of the code C, namely

$f(x) \in \mathbb{F}_q[x]$ and $\deg(f(x)) \leq k-1$. The first condition can be expressed in terms of valuations, namely $\nu_{p(x)}(f(x)) \geq 0$ for all valuations $\nu_{p(x)}$ of $\mathbb{F}_q(x)$. The second requirement can be formulated in terms of the valuation ν_∞ of $\mathbb{F}_q(x)$ that is defined as follows: $\nu_\infty(0) = \infty$ and $\nu_\infty(f(x)) = -\deg(f(x))$ for a nonzero $f(x) \in \mathbb{F}_q(x)$. Then $\deg(f(x)) \leq k-1$ is of course equivalent to $\nu_\infty(f(x)) \geq -(k-1)$. The upshot of this discussion is that the code C can be completely described by means of the language of valuations.

Now we proceed from rational function fields to more general global function fields and we formally define valuations. A *global function field* F over $\mathbb{F}_q$ is a finite extension (in the sense of field theory) of the rational function field $\mathbb{F}_q(x)$. The step from $\mathbb{F}_q(x)$ to F is analogous to the step from the field $\mathbb{Q}$ of rational numbers to an algebraic number field. We call $\mathbb{F}_q$ the *full constant field* of F if every element of F which is a root of a nonzero polynomial over $\mathbb{F}_q$ is actually in $\mathbb{F}_q$. We write $F/\mathbb{F}_q$ to signify that F is a global function field with full constant field $\mathbb{F}_q$. A *valuation* ν of F is a map $\nu : F \to \mathbb{R} \cup \{\infty\}$ satisfying the following axioms: (i) $\nu(f) = \infty$ if and only if $f = 0$; (ii) $\nu(fh) = \nu(f) + \nu(h)$ for all $f, h \in F$; (iii) $\nu(f+h) \geq \min(\nu(f), \nu(h))$ for all $f, h \in F$; (iv) the image of ν contains more than the two elements 0 and ∞. It is a simple consequence of the axioms that $\nu(c) = 0$ for all $c \in \mathbb{F}_q^*$. Furthermore, it is easily verified that the maps $\nu_{p(x)}$ and ν_∞ in the previous paragraph are indeed valuations of $\mathbb{F}_q(x)$. A *place* of the global function field F is an equivalence class of valuations of F, where two valuations of F are considered equivalent if one is obtained from the other by multiplying by a positive constant. Every place P of F contains a uniquely determined normalized valuation ν_P, that is, the image of the map ν_P is $\mathbb{Z} \cup \{\infty\}$. Let $\mathbb{P}_F$ denote the set of all places of F. For $F = \mathbb{F}_q(x)$ there is a one-to-one correspondence between $\mathbb{P}_F$ and the set $\{p(x) \in \mathbb{F}_q[x] : p(x) \text{ monic irreducible}\} \cup \{\infty\}$; in other words, the normalized valuations of $\mathbb{F}_q(x)$ are exactly given by $\nu_{p(x)}$ with $p(x) \in \mathbb{F}_q[x]$ monic irreducible and ν_∞.

For a global function field $F/\mathbb{F}_q$ and a place $P \in \mathbb{P}_F$, we introduce its valuation ring $O_P = \{f \in F : \nu_P(f) \geq 0\}$, its unique maximal ideal $M_P = \{f \in F : \nu_P(f) \geq 1\}$, and its residue class field O_P/M_P which can be identified with a finite extension field of $\mathbb{F}_q$. The degree of this extension is called the *degree* of the place P and denoted by $\deg(P)$. If $\deg(P) = 1$, then P is called a *rational place*. For a rational place P of $F/\mathbb{F}_q$ and every $f \in O_P$, the residue class of f modulo M_P can be identified with a unique element of $\mathbb{F}_q$ which is denoted by $f(P)$. A *divisor* D of F is a formal sum

$$D = \sum_{P \in \mathbb{P}_F} z_P \, P$$

with coefficients $z_P \in \mathbb{Z}$ for all $P \in \mathbb{P}_F$ and all but finitely many $z_P = 0$. Divisors of F can be added by adding corresponding coefficients. We write $D \geq 0$ if $z_P \geq 0$

for all $P \in \mathbb{P}_F$. The *degree* $\deg(D)$ of D is defined by

$$\deg(D) = \sum_{P \in \mathbb{P}_F} z_P \deg(P).$$

The *principal divisor* $\operatorname{div}(f)$ of $f \in \mathbb{F}^*$ is given by

$$\operatorname{div}(f) = \sum_{P \in \mathbb{P}_F} \nu_P(f)\, P.$$

The *Riemann-Roch space*

$$\mathcal{L}(D) = \{f \in F^* : \operatorname{div}(f) + D \geq 0\} \cup \{0\}$$

associated with the divisor D is a finite-dimensional vector space over $\mathbb{F}_q$. Much more information on global function fields can be found in the books [148] and [192].

We have now assembled all the tools that are needed for the introduction of algebraic-geometry codes. First we return one more time to the special case of the code C in Remark 3.5.34 with $a_j = 1 \in \mathbb{F}_q$ for $1 \leq j \leq n$. Let P_∞ be the place of $\mathbb{F}_q(x)$ containing the normalized valuation ν_∞. Consider the divisor $D = (k-1)P_\infty$ of $\mathbb{F}_q(x)$. Then by an earlier discussion it is clear that $f \in \mathbb{F}_q[x]_{<k}$ if and only if $f \in \mathcal{L}(D)$. For $1 \leq j \leq n$, let P_j be the place of $\mathbb{F}_q(x)$ corresponding to the monic irreducible polynomial $x - c_j \in \mathbb{F}_q[x]$. As we have seen earlier, for $f \in \mathcal{L}(D)$ we have $f - f(c_j) \in M_{P_j}$, and so $f(P_j) = f(c_j)$. Therefore the code C can be described as

$$C = \{(f(P_1), \ldots, f(P_n)) \in \mathbb{F}_q^n : f \in \mathcal{L}(D)\}.$$

It is now pretty obvious how to generalize this construction. Let n be the length of the code to be constructed and let $F/\mathbb{F}_q$ be a global function field with at least n distinct rational places. Choose distinct rational places $P_1, \ldots, P_n$ of F and a divisor $D = \sum_{P \in \mathbb{P}_F} z_P\, P$ of F with $z_{P_j} = 0$ for $1 \leq j \leq n$. Then

$$C(P_1, \ldots, P_n; D) := \{(f(P_1), \ldots, f(P_n)) \in \mathbb{F}_q^n : f \in \mathcal{L}(D)\}$$

is an *algebraic-geometry code*. It is easily seen to be a subspace of $\mathbb{F}_q^n$. In order to guarantee that it has a positive dimension (and thus is a linear code in the sense of Definition 3.2.10), conditions on the divisor D are needed. Here the genus g of F, a nonnegative integer g depending only on F, is involved. If now $g \leq \deg(D) < n$, then $C(P_1, \ldots, P_n; D)$ has dimension $k \geq \deg(D)+1-g$. Its minimum distance d satisfies $d \geq n - \deg(D)$. We refer to the books [148] and [192] for detailed treatments of algebraic-geometry codes. You may wonder why we speak of an "algebraic-geometry code" and not of a "global-function-field code". This has historical reasons: the first constructions of algebraic-geometry codes used the theory of algebraic curves over finite fields which belongs to algebraic geometry. As a matter of fact, the theory of algebraic curves over finite fields has close links to the theory of global function fields (for an in-depth explanation of this connection see [148, Chapter 3]). Consequently, algebraic-geometry codes can be completely described via global function fields.

Algebraic-geometry codes have important implications for the asymptotic theory of codes. For every prime power q and every integer $g \geq 0$, let $N_q(g)$ be the maximum number of rational places that a global function field $F/\mathbb{F}_q$ of genus g can have. Furthermore, we put

$$A(q) = \limsup_{g \to \infty} \frac{N_q(g)}{g}.$$

It is known that $0 < A(q) \leq q^{1/2} - 1$ for all q and that $A(q) = q^{1/2} - 1$ if q is a square (see [147, Chapter 5]). By using algebraic-geometry codes, we get a lower bound on the function α_q in (3.22) for all prime powers q, namely

$$\alpha_q(\delta) \geq 1 - \frac{1}{A(q)} - \delta \qquad \text{for } 0 < \delta < \frac{q-1}{q}. \tag{3.23}$$

A comparison of the right-hand sides of (3.22) and (3.23) shows that, at least for squares $q \geq 49$, the lower bound in (3.23) is larger than the lower bound in (3.22) for all δ in a subinterval of $(0, (q-1)/q)$ containing the number $(q-1)/(2q-1)$. If one considers arbitrary (hence also nonlinear) codes, then the positive constant $\log_q(1+q^{-3})$ can be added on the right-hand side of (3.23). For a proof of this result and of the bound (3.23), we refer to [148, Section 5.3].

We already encountered character sums for finite fields in Subsection 3.3.6 on irreducible cyclic codes. There are quite a number of other fascinating applications of character sums to coding theory. These concern mainly BCH codes, the dual codes of BCH codes, and the theory of perfect codes. A nice survey of applications of character sums to coding theory is presented in [162, Chapter 13].

It is a remarkable fact that codes can be used for the construction of cryptographic schemes. Historically the first code-based cryptographic scheme was the *McEliece cryptosystem*, a public-key cryptosystem that still remains unbroken in its general form. As usual in cryptography, we describe the scheme from the perspective of two users Alice and Bob. Let C be a linear $[n, k, d]$ code over $\mathbb{F}_q$ and let G be a generator matrix of C. The matrix G is part of the private key of Bob. Next, Bob chooses two more matrices over $\mathbb{F}_q$, namely a nonsingular $k \times k$ matrix N and an $n \times n$ matrix Q that is obtained from a nonsingular $n \times n$ diagonal matrix by arbitrary row permutations. The matrices G, N, and Q form Bob's private key. The public key of Bob is the $k \times n$ matrix $G' = NGQ$ which may be viewed as a scrambled version of G. The admissible plaintexts in the McEliece cryptosystem are vectors $\mathbf{u} \in \mathbb{F}_q^k$. If Alice wants to encrypt the plaintext $\mathbf{u} \in \mathbb{F}_q^k$ destined for Bob, she chooses a random vector $\mathbf{v} \in \mathbb{F}_q^n$ with Hamming weight $w(\mathbf{v}) \leq t := \lfloor (d-1)/2 \rfloor$ and uses Bob's public key G' to compute the ciphertext $\mathbf{y} = \mathbf{u}G' + \mathbf{v} \in \mathbb{F}_q^n$. If Bob receives the ciphertext $\mathbf{y}$, he first computes $\mathbf{y}' = \mathbf{y}Q^{-1} = \mathbf{u}NG + \mathbf{v}Q^{-1}$. Now $w(\mathbf{y}' - \mathbf{u}NG) = w(\mathbf{v}Q^{-1}) = w(\mathbf{v}) \leq t$ and $\mathbf{u}NG = (\mathbf{u}N)G$ is a codeword in C.

Therefore $\mathbf{y}'$ is like a received word that can be corrected by the code C to produce the original word $\mathbf{u}N$. From this, Bob recovers the plaintext $\mathbf{u} = (\mathbf{u}N)N^{-1}$.

A related public-key cryptosystem is the *Niederreiter cryptosystem*. Given the linear code C as above, Bob chooses a parity-check matrix H of C as part of his private key. Furthermore, Bob selects a matrix Q as in the McEliece cryptosystem as well as a nonsingular $(n-k) \times (n-k)$ matrix M over $\mathbb{F}_q$. The matrices H, M, and Q form Bob's private key, whereas his public key is the $(n-k) \times n$ matrix $H' = MHQ$ which may be regarded as a scrambled version of H. The admissible plaintexts in the Niederreiter cryptosystem are column vectors $\mathbf{x}^\top \in \mathbb{F}_q^n$ with Hamming weight at most $t := \lfloor (d-1)/2 \rfloor$. Alice encrypts $\mathbf{x}^\top$ by computing the ciphertext $\mathbf{z}^\top = H'\mathbf{x}^\top$ using Bob's public key H'. The decryption proceeds again by using a decoding algorithm for C. With corresponding choices of code parameters, the McEliece and Niederreiter cryptosystems have basically equivalent security levels (see [148, Theorem 6.4.1]). The Niederreiter cryptosystem has the advantage that a digital signature scheme (called the CFS scheme) can be derived from it. Detailed discussions of the McEliece and Niederreiter cryptosystems and of various other code-based cryptographic schemes can be found in the book [8] and in the survey article [156]. It turns out that certain quasicyclic codes (see the first paragraph of this section) are eminently suitable for code-based cryptography (see again [8]).

Exercises

3.1 Prove that a code C with minimum distance $d(C)$ cannot correct more than $\lfloor (d(C)-1)/2 \rfloor$ errors in general.

3.2 Prove that a code C with minimum distance $d(C)$ cannot detect more than $d(C)-1$ errors in general.

3.3 Consider the binary code C of length 4 consisting of the codewords

$$\mathbf{c}_1 = (0,0,0,0), \quad \mathbf{c}_2 = (0,0,0,1), \quad \mathbf{c}_3 = (0,0,1,1),$$
$$\mathbf{c}_4 = (1,0,0,0), \quad \mathbf{c}_5 = (1,0,0,1), \quad \mathbf{c}_6 = (1,1,0,0).$$

Suppose that the word $\mathbf{v} = (1,1,1,0) \in \mathbb{F}_2^4$ is received. Use nearest neighbor decoding to determine the most likely codeword in C that was sent.

3.4 Let V be a vector space over $\mathbb{F}_q$. Prove that a nonempty subset W of V is a subspace of V if and only if $c\mathbf{u} + \mathbf{w} \in W$ for all $\mathbf{u}, \mathbf{w} \in W$ and all $c \in \mathbb{F}_q$.

3.5 Prove that if V is a vector space over $\mathbb{F}_q$, then the intersection of any collection of subspaces of V is a subspace of V.

3.6 Prove that if V_1 and V_2 are subspaces of $\mathbb{F}_q^n$, then $V_1 + V_2 := \{\mathbf{v}_1 + \mathbf{v}_2 : \mathbf{v}_1 \in V_1,\ \mathbf{v}_2 \in V_2\}$ is also a subspace of $\mathbb{F}_q^n$.

3.7 Prove that if V_1 and V_2 are subspaces of $\mathbb{F}_q^n$, then $(V_1 + V_2)^\perp = V_1^\perp \cap V_2^\perp$.

3.8 Prove that if V_1 and V_2 are subspaces of $\mathbb{F}_q^n$, then $(V_1 \cap V_2)^\perp = V_1^\perp + V_2^\perp$.

3.9 Prove in detail that multiplication of matrices over finite fields is associative.

3.10 Prove in detail that $A(B + C) = AB + AC$ and $(A + B)C = AC + BC$ for matrices A, B, and C over $\mathbb{F}_q$ of compatible sizes.

3.11 Prove that equivalent linear codes have the same parameters n, k, and d.

3.12 Prove that if two nontrivial linear codes C_1 and C_2 are equivalent, then their dual codes $C_1^\perp$ and $C_2^\perp$ are equivalent.

3.13 Prove that for a nontrivial linear code C, a parity-check matrix of $C^\perp$ is given by a generator matrix of C.

3.14 Consider the binary linear code C with generator matrix

$$G = \begin{pmatrix} 1\,0\,0\,1\,1\,0 \\ 0\,1\,0\,0\,1\,1 \\ 0\,0\,1\,1\,1\,1 \end{pmatrix}.$$

(a) Determine a parity-check matrix of C.

(b) Set up a table of coset leaders and syndromes for C. (Note that here not every coset has a unique coset leader, so for cosets with several possible coset leaders you can make an arbitrary choice among the candidate coset leaders.)

(c) Suppose that the received word is $\mathbf{v} = (1\ 1\ 0\ 0\ 1\ 0) \in \mathbb{F}_2^6$. Use the syndrome decoding algorithm to determine the most likely codeword in C that was sent.

3.15 Consider the binary linear code C with generator matrix

$$G = \begin{pmatrix} 1\,1\,0\,0\,1 \\ 0\,1\,1\,1\,0 \\ 1\,0\,1\,0\,0 \end{pmatrix}.$$

Determine the weight enumerators of C and $C^\perp$ and check the MacWilliams identity.

3.16 Prove that equivalent linear codes have the same weight enumerator.

3.17 Prove that every nonzero subspace of a self-orthogonal code is self-orthogonal.

3.18 Let C be a ternary self-orthogonal code. Prove that the Hamming weight of each codeword in C is divisible by 3.

3.19 For an odd integer $n \geq 3$, let C be a binary self-orthogonal $[n, (n-1)/2]$ code. Prove that a generator matrix of the dual code $C^{\perp}$ is obtained from a generator matrix of C by appending as a new row vector the all-one vector $(1, \ldots, 1) \in \mathbb{F}_2^n$.

3.20 Prove that if a binary linear code C has the property that the Hamming weight of each codeword in C is divisible by 4, then C is self-orthogonal.

3.21 Find a formula for the number of cyclic codes over $\mathbb{F}_q$ of given length n in terms of the canonical factorization of $x^n - 1 \in \mathbb{F}_q[x]$ over $\mathbb{F}_q$.

3.22 For cyclic codes C_1 and C_2 over $\mathbb{F}_q$ of the same length and with generator polynomials $g_1(x)$ and $g_2(x)$, respectively, prove that $C_1 \subseteq C_2$ if and only if $g_2(x)$ divides $g_1(x)$ in $\mathbb{F}_q[x]$.

3.23 Let $x^n - 1 = g(x)h(x)$ in $\mathbb{F}_q[x]$ with $\deg(g(x)) \geq 1$ and $\deg(h(x)) \geq 1$. Prove that the cyclic code over $\mathbb{F}_q$ with generator polynomial $g(x)$ is self-orthogonal if and only if the reciprocal polynomial of $h(x)$ divides $g(x)$ in $\mathbb{F}_q[x]$.

3.24 Given the binary cyclic code C of length 6 with generator polynomial $g(x) = x^4 + x^3 + x + 1 \in \mathbb{F}_2[x]$, determine a generator matrix of C and a parity-check matrix of C.

3.25 Let C be the binary cyclic code of length 7 with generator polynomial $g(x) = x^3 + x + 1 \in \mathbb{F}_2[x]$.

(a) Determine the minimum distance of C.

(b) Decode the received word $(0\ 1\ 1\ 1\ 1\ 1\ 0) \in \mathbb{F}_2^7$ with the code C.

3.26 Let C be the binary cyclic code of length 15 with generator polynomial $g(x) = x^8 + x^7 + x^6 + x^4 + 1 \in \mathbb{F}_2[x]$.

(a) Determine the minimum distance of C, for instance by considering a parity-check matrix of C.

(b) Decode the received word

$$(1\ 1\ 0\ 0\ 1\ 1\ 1\ 0\ 1\ 1\ 0\ 0\ 0\ 1\ 0) \in \mathbb{F}_2^{15}$$

with the code C by using Algorithm 3.3.49.

3.27 Prove that there is no cyclic self-dual code over $\mathbb{F}_q$ when q is odd.

3.28 Prove that the Mattson-Solomon polynomial of the all-one vector $(1, \ldots, 1) \in \mathbb{F}_q^n$ is the constant polynomial n, where n is considered as an element of the prime subfield of $\mathbb{F}_q$.

3.29 Let $M_{\mathbf{v}}(x)$ be the Mattson-Solomon polynomial of $\mathbf{v} \in \mathbb{F}_q^n$, where $n \geq 2$ and $\gcd(n, q) = 1$, and let γ be a primitive nth root of unity in a finite extension field of $\mathbb{F}_q$. Prove that for every integer t with $0 \leq t \leq n-1$, the Mattson-Solomon polynomial of the cyclic shift $\mathbf{v}^t$ is given by $M_{\mathbf{v}}(\gamma^{-t}x)$.

3.30 Generalize the hypotheses in Theorem 3.3.34 by assuming that for some $r \in \mathbb{N}$ with $\gcd(r, n) = 1$ we have $g(\gamma^{b+ir}) = 0$ for $0 \leq i \leq d-2$. Show again that the minimum distance of C is at least d.

3.31 Show by an example that a code equivalent to a cyclic code need not be cyclic.

3.32 Prove that the following is the complete list of binary MDS codes: (i) $C_1 = \mathbb{F}_2^n$ for $n \geq 1$; (ii) $C_2 = \{\mathbf{0}, \mathbf{1}\}$ with $\mathbf{0} \in \mathbb{F}_2^n$ and the all-one vector $\mathbf{1} = (1, \ldots, 1) \in \mathbb{F}_2^n$ for $n \geq 1$; (iii) every code C_3 equivalent to $C_2^{\perp}$ for $n \geq 2$. (Hint: consider generator matrices in standard form.)

3.33 Prove the following version of the Gilbert-Varshamov bound: if n, k, and d are integers with $1 \leq k \leq n$, $1 \leq d \leq n$, and

$$\sum_{i=0}^{d-1} \binom{n}{i} (q-1)^i < q^{n-k+1},$$

then there exists a linear $[n, k, d]$ code over $\mathbb{F}_q$.

3.34 Decode the received word $(1\ 1\ 0\ 0\ 0\ 0\ 1) \in \mathbb{F}_2^7$ with the Hamming code $\mathrm{Ham}(3, 2)$.

3.35 Prove that $G_{23}^{\perp} \subseteq G_{23}$, or in other words that $G_{23}^{\perp}$ is self-orthogonal. (Hint: consider the generator polynomials of these cyclic codes.)

3.36 Prove that $G_{11}^{\perp} \subseteq G_{11}$, or in other words that $G_{11}^{\perp}$ is self-orthogonal. (Hint: consider the generator polynomials of these cyclic codes.)

3.37 Let $\alpha \in \mathbb{F}_{16}$ be a root of $x^4 + x + 1 \in \mathbb{F}_2[x]$. Find the minimal polynomial of α^{11} over $\mathbb{F}_2$.

3.38 Determine the generator polynomial of a Reed-Solomon code over $\mathbb{F}_{16}$ of dimension 11 and find a parity-check matrix of such a code.

3.39 Prove that the dual code of a Reed-Solomon code is again a Reed-Solomon code.

3.40 Determine the generator polynomials of all binary BCH codes of length 31 and designed distance 5.

Chapter 4

Quasi-Monte Carlo Methods

Good lattice points and nets
are so much better bets
in tough numerical integration
since they beat stochastic simulation
hands down and in straight sets.

4.1 Numerical integration and uniform distribution

4.1.1 The one-dimensional case

There are many scientific as well as real-world applications where we run into the problem of computing a definite integral. In calculus courses you are taught that a definite integral $\int_a^b f(u)du$ is evaluated by the fundamental theorem of integral calculus which says that

$$\int_a^b f(u)du = F(b) - F(a), \tag{4.1}$$

where the function F is an antiderivative of the integrand f. What you are often not told is that there are many cases where F cannot be expressed in finite terms by means of elementary functions, and in such situations the formula (4.1) is useless for computational purposes. Examples are $\int_0^1 e^{-u^2}du$ and $\int_0^1 (\sin u)(u+1)^{-1}du$. We then have to settle for numerical approximations of $\int_a^b f(u)du$. The process of approximately computing definite integrals with a sufficient degree of precision is called *numerical integration*.

We start with the one-dimensional case, that is, the case considered in (4.1) where the integrand f is a real-valued function of a single variable u. One-dimensional numerical integration is an area of numerical analysis with a long tradition, and indeed some very effective one-dimensional numerical integration techniques are known for several centuries. Classical numerical integration rules, such as the midpoint rule, the trapezoidal rule, and Simpson's rule, are based on approximations of the form

$$\int_a^b f(u)du \approx \sum_{n=1}^{N} w_n f(x_n),$$

where the *integration nodes* (or simply *nodes*) $x_1, \dots, x_N$ are points lying in the integration domain $[a,b]$ and the coefficients $w_1, \dots, w_N$ are "weights" associated with these points. It is usually assumed that $\sum_{n=1}^{N} w_n = b - a$ since this condition guarantees that at least constant functions f are integrated correctly by the numerical integration scheme. Particularly simple and attractive rules are *equal-weight rules* where $w_n = (b-a)/N$ for $1 \le n \le N$. We assume in the following that the integration domain is the unit interval $[0,1]$; this is achieved by a simple change of variable.

An equal-weight rule for the interval $[0,1]$ has the form

$$\int_0^1 f(u)du \approx \frac{1}{N}\sum_{n=1}^{N} f(x_n) \tag{4.2}$$

with nodes $x_1, \dots, x_N \in [0,1]$. A prominent equal-weight rule is the already mentioned *midpoint rule* which is given by

$$\int_0^1 f(u)du \approx \frac{1}{N}\sum_{n=1}^{N} f\big(\frac{2n-1}{2N}\big). \tag{4.3}$$

The name stems from a way of interpreting this rule, namely that we split up $[0,1]$ into the N subintervals $[0,1/N], [1/N,2/N], \dots, [(N-1)/N,1]$, and then we take the midpoint of each subinterval as a node.

It is essential for practical computations that every numerical integration rule be accompanied by an upper bound on the error that is committed by the approximation to the given definite integral. As a simple illustration, we present an error bound for the midpoint rule under a smoothness condition on the integrand.

Proposition 4.1.1 *Let f be a real-valued function on $[0,1]$ which has a continuous second derivative f'' on $[0,1]$. Then for every integer $N \ge 1$,*

$$\left|\int_0^1 f(u)du - \frac{1}{N}\sum_{n=1}^{N} f\big(\frac{2n-1}{2N}\big)\right| \le \frac{1}{24N^2}\max_{0\le u\le 1}|f''(u)|. \tag{4.4}$$

Proof. We write

$$\int_0^1 f(u)du - \frac{1}{N}\sum_{n=1}^{N} f\big(\frac{2n-1}{2N}\big) = \sum_{n=1}^{N}\int_{(n-1)/N}^{n/N}\Big(f(u) - f\big(\frac{2n-1}{2N}\big)\Big)du. \tag{4.5}$$

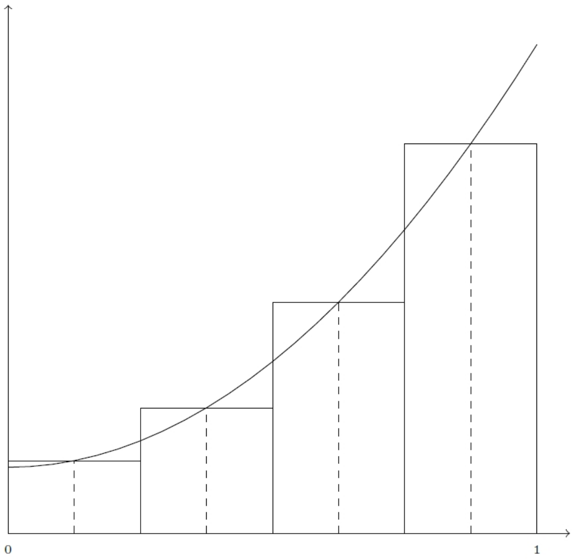

Figure 4.1: The midpoint rule

Since $\int_{(n-1)/N}^{n/N}(u-\frac{2n-1}{2N})du=0$ for $1\le n\le N$, we obtain

$$\int_{(n-1)/N}^{n/N}\Big(f(u)-f\big(\frac{2n-1}{2N}\big)\Big)du=\int_{(n-1)/N}^{n/N}\Big(f(u)-f\big(\frac{2n-1}{2N}\big)-f'\big(\frac{2n-1}{2N}\big)\big(u-\frac{2n-1}{2N}\big)\Big)du,$$

and so

$$\left|\int_{(n-1)/N}^{n/N}\Big(f(u)-f\big(\frac{2n-1}{2N}\big)\Big)du\right|\le\int_{(n-1)/N}^{n/N}\Big|f(u)-f\big(\frac{2n-1}{2N}\big)-f'\big(\frac{2n-1}{2N}\big)\big(u-\frac{2n-1}{2N}\big)\Big|du.$$

By Taylor's theorem,

$$\Big|f(u)-f\big(\frac{2n-1}{2N}\big)-f'\big(\frac{2n-1}{2N}\big)\big(u-\frac{2n-1}{2N}\big)\Big|\le\frac{m}{2}\big(u-\frac{2n-1}{2N}\big)^2\qquad\text{for } 0\le u\le 1$$

with $m=\max_{0\le u\le 1}|f''(u)|$. It follows that

$$\left|\int_{(n-1)/N}^{n/N}\Big(f(u)-f\big(\frac{2n-1}{2N}\big)\Big)du\right|\le\frac{m}{2}\int_{(n-1)/N}^{n/N}\big(u-\frac{2n-1}{2N}\big)^2=\frac{m}{24N^3}.$$

Summing over $n=1,\dots,N$ and taking into account (4.5), we arrive at the desired bound. □

Remark 4.1.2 The error bound in (4.4) is in general best possible, in the sense that we can have equality in (4.4). Just take $f(u)=u^2$ on $[0,1]$, then a straightforward calculation shows that

$$\int_0^1 f(u)du-\frac{1}{N}\sum_{n=1}^{N}f\big(\frac{2n-1}{2N}\big)=\frac{1}{12N^2},$$

which agrees with the right-hand side of (4.4).

For a fixed integrand f satisfying the smoothness condition in Proposition 4.1.1, the error bound in (4.4) becomes smaller as the number N of nodes increases. Moreover, the error bound tends to 0 as $N \to \infty$. We express the latter fact by saying that the midpoint rule converges. Any reasonable numerical integration scheme should have this property. We will not pursue classical numerical integration rules any further since we want to focus on the applications of number theory to numerical integration. We refer to the standard monograph by Davis and Rabinowitz [35] and to the more recent book by Brass and Petras [14] for a detailed coverage of classical numerical integration rules.

Now we return to the general equal-weight rule (4.2). We try to obtain a convergent numerical integration scheme by constructing a sequence $x_1, x_2, \ldots$ of points in $[0,1]$ such that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(x_n) = \int_0^1 f(u)du$$

for a given integrand f. In fact, there are sequences $x_1, x_2, \ldots$ for which this limit relation holds not only for a single integrand f, but for a wide family of integrands. Such sequences are called "uniformly distributed" in number theory. For the formal definition of a uniformly distributed sequence, we work with the family of Riemann-integrable functions. It is customary to consider sequences of points from the half-open interval $[0,1)$ since all classical constructions of uniformly distributed sequences produce points from this interval.

Definition 4.1.3 A sequence $x_1, x_2, \ldots$ of points in the interval $[0,1)$ is *uniformly distributed* (in $[0,1)$) if

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(x_n) = \int_0^1 f(u)du \tag{4.6}$$

for every real-valued Riemann-integrable function f on $[0,1]$.

Remark 4.1.4 If you know Lebesgue integrals, then you will understand that the limit relation (4.6) cannot hold for all Lebesgue-integrable functions on $[0,1]$. Let $x_1, x_2, \ldots$ be any sequence of points in $[0,1)$ and consider the set $S = \{x_1, x_2, \ldots\} \subset [0,1]$. Then with f being the characteristic function of S, that is, $f(u) = 1$ if $u \in S$ and $f(u) = 0$ if $u \in [0,1] \setminus S$, it is trivial that the left-hand side of (4.6) is equal to 1, whereas the right-hand side of (4.6) is equal to 0.

There are various other characterizations of uniformly distributed sequences in $[0,1)$ that use different families of functions for which we require the validity of the limit relation (4.6). The following approximation principle is convenient in this context.

Lemma 4.1.5 *Let $x_1, x_2, \dots$ be a sequence of points in $[0,1)$ and let $\mathcal{G}$ be a nonempty family of real-valued Riemann-integrable functions on $[0,1]$ such that*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} g(x_n) = \int_0^1 g(u)du \qquad \text{for all } g \in \mathcal{G}.$$

Let f be a real-valued Riemann-integrable function on $[0,1]$ such that for every $\varepsilon > 0$ there exist functions $g_{1,\varepsilon}, g_{2,\varepsilon} \in \mathcal{G}$ with $g_{1,\varepsilon}(u) \le f(u) \le g_{2,\varepsilon}(u)$ for all $u \in [0,1]$ and

$$\int_0^1 (g_{2,\varepsilon}(u) - g_{1,\varepsilon}(u))du \le \varepsilon.$$

Then

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(x_n) = \int_0^1 f(u)du.$$

Proof. For every $\varepsilon > 0$ we obtain the chain of inequalities and identities

$$\begin{aligned} \int_0^1 f(u)du - \varepsilon \le \int_0^1 g_{1,\varepsilon}(u)du &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} g_{1,\varepsilon}(x_n) \\ &\le \liminf_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(x_n) \le \limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(x_n) \\ &\le \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} g_{2,\varepsilon}(x_n) = \int_0^1 g_{2,\varepsilon}(u)du \le \int_0^1 f(u)du + \varepsilon. \end{aligned}$$

Letting $\varepsilon \to 0+$, we get the desired result. □

For a subinterval J of $[0,1]$, we write c_J for the characteristic function of J, that is, $c_J(u) = 1$ if $u \in J$ and $c_J(u) = 0$ if $u \in [0,1] \setminus J$.

Theorem 4.1.6 *A sequence $x_1, x_2, \dots$ of points in $[0,1)$ is uniformly distributed if and only if*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} c_J(x_n) = \lambda(J) \tag{4.7}$$

for every subinterval J of $[0,1]$, where $\lambda(J)$ is the length of the interval J.

Proof. The necessity is trivial by (4.6) since c_J is Riemann-integrable and

$$\int_0^1 c_J(u)du = \lambda(J).$$

In order to prove the sufficiency, we note that, by linearity, the limit relation (4.6) holds for all real-valued step functions on $[0,1]$ (that is, for all finite $\mathbb{R}$-linear combinations of characteristic functions of subintervals of $[0,1]$). Let $\mathcal{G}$ be the family of all real-valued step functions on $[0,1]$. Then a given real-valued Riemann-integrable function f on $[0,1]$ satisfies the condition in Lemma 4.1.5 by the definition of the Riemann integral, and so an application of this lemma completes the proof. □

Theorem 4.1.7 *A sequence $x_1, x_2, \ldots$ of points in $[0,1)$ is uniformly distributed if and only if*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(x_n) = \int_0^1 f(u)du$$

for every real-valued continuous function f on $[0,1]$.

Proof. The necessity is trivial since every real-valued continuous function on $[0,1]$ is Riemann-integrable. In order to prove the sufficiency, we show (4.7) for every subinterval J of $[0,1]$. We assume in fact that $J=[a,b]$ with $0<a<b<1$; the remaining case is treated with obvious modifications. Let $\mathcal{G}$ be the family of all real-valued continuous functions on $[0,1]$. In view of Lemma 4.1.5, it suffices to construct, for $0<\varepsilon<\min(2a, 2-2b, b-a)$, two functions $g_{1,\varepsilon}, g_{2,\varepsilon} \in \mathcal{G}$ such that the condition in Lemma 4.1.5 is satisfied for $f=c_J$. Let $g_{1,\varepsilon}$ be the piecewise linear continuous function which agrees with c_J on the set $[0,a) \cup [a+\varepsilon/2, b-\varepsilon/2] \cup (b,1]$; on the interval $[a, a+\varepsilon/2]$ the graph of $g_{1,\varepsilon}$ is the line segment connecting the points $(a,0)$ and $(a+\varepsilon/2, 1)$ in $\mathbb{R}^2$; on the interval $[b-\varepsilon/2, b]$ the graph of $g_{1,\varepsilon}$ is the line segment connecting the points $(b-\varepsilon/2, 1)$ and $(b,0)$ in $\mathbb{R}^2$. Let $g_{2,\varepsilon}$ be the piecewise linear continuous function which agrees with c_J on the set $[0, a-\varepsilon/2) \cup [a,b] \cup (b+\varepsilon/2, 1]$; on the interval $[a-\varepsilon/2, a]$ the graph of $g_{2,\varepsilon}$ is the line segment connecting the points $(a-\varepsilon/2, 0)$ and $(a,1)$ in $\mathbb{R}^2$; on the interval $[b, b+\varepsilon/2]$ the graph of $g_{2,\varepsilon}$ is the line segment connecting the points $(b,1)$ and $(b+\varepsilon/2, 0)$ in $\mathbb{R}^2$ (see Figure 4.2).

Then $g_{1,\varepsilon}(u) \le c_J(u) \le g_{2,\varepsilon}(u)$ for all $u \in [0,1]$ and

$$\int_0^1 (g_{2,\varepsilon}(u) - g_{1,\varepsilon}(u))du = \varepsilon.$$

Thus, we have obtained suitable functions $g_{1,\varepsilon}, g_{2,\varepsilon} \in \mathcal{G}$. □

In number theory one often arrives at the situation where a sequence of points in $[0,1)$ is obtained by taking fractional parts in a sequence of real numbers. The *fractional part* $\{x\}$ of a real number x is defined by $\{x\} = x - \lfloor x \rfloor$. The following definition refers to this situation.

Definition 4.1.8 A sequence $x_1, x_2, \ldots$ of real numbers is *uniformly distributed modulo* 1 if the sequence $\{x_1\}, \{x_2\}, \ldots$ of fractional parts is uniformly distributed in $[0,1)$.

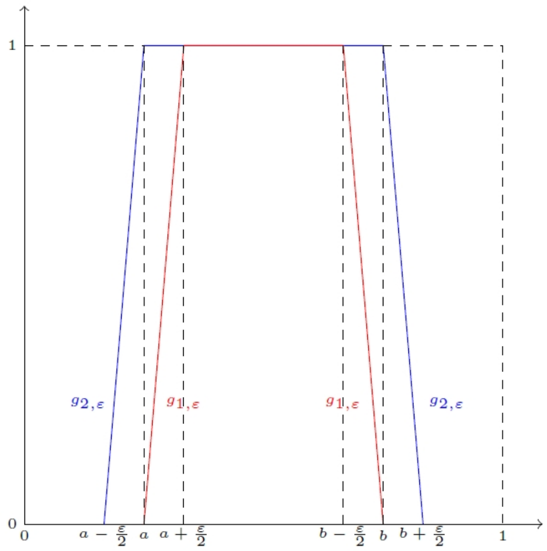

Figure 4.2: The graphs of $g_{1,\varepsilon}$ and $g_{2,\varepsilon}$

There is a famous criterion for uniform distribution modulo 1, the Weyl criterion, which goes back all the way to the celebrated paper of Weyl [200] from 1916 in which he introduced the general theory of uniformly distributed sequences. Hermann Weyl (1885–1955) later moved on to even bigger things, doing fundamental work in functional analysis, differential geometry, and mathematical physics.

Theorem 4.1.9 (Weyl Criterion) *The sequence $x_1, x_2, \ldots$ of real numbers is uniformly distributed modulo* 1 *if and only if*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} e^{2\pi \mathrm{i} h x_n} = 0 \qquad \textit{for all } h \in \mathbb{N}. \tag{4.8}$$

Proof. Let the sequence $x_1, x_2, \ldots$ be uniformly distributed modulo 1 and consider the function $f(u) = \cos 2\pi h u$ on $\mathbb{R}$ for a fixed $h \in \mathbb{N}$. Then by Definition 4.1.3,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \cos 2\pi h x_n = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \cos 2\pi h \{x_n\} = \int_0^1 \cos 2\pi h u \, du = 0.$$

Similarly, we obtain

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \sin 2\pi h x_n = 0,$$

and the fundamental identity $e^{2\pi \mathrm{i} h u} = \cos 2\pi h u + \mathrm{i} \sin 2\pi h u$ for all $u \in \mathbb{R}$ yields the limit relation in (4.8).

Conversely, suppose that the sequence $x_1, x_2, \ldots$ satisfies (4.8). Since $e^{-2\pi \mathrm{i} h u}$ is the complex conjugate of $e^{2\pi \mathrm{i} h u}$, the limit relation in (4.8) holds also for all negative integers h. Let $\varepsilon > 0$ be given and let f be any one of the two functions $g_{1,\varepsilon}$ and $g_{2,\varepsilon}$ in the proof of Theorem 4.1.7. Then $f(0) = f(1) = 0$, and so f can be extended to a real-valued continuous function on $\mathbb{R}$ with period 1. Hence by the Weierstrass approximation theorem, for every $\delta > 0$ there exists a trigonometric polynomial $\Psi_\delta(u)$, that is, a finite linear combination of functions of the type $e^{2\pi i h u}$, $h \in \mathbb{Z}$, with complex coefficients, such that

$$\max_{u\in\mathbb{R}} |f(u) - \Psi_\delta(u)| \leq \delta. \tag{4.9}$$

Now for every positive integer N,

$$\begin{aligned}\Big|\int_0^1 f(u)du - \frac{1}{N}\sum_{n=1}^N f(x_n)\Big| \leq{}& \Big|\int_0^1 (f(u) - \Psi_\delta(u))\, du\Big| \\ &+\Big|\int_0^1 \Psi_\delta(u)du - \frac{1}{N}\sum_{n=1}^N \Psi_\delta(x_n)\Big| \\ &+\Big|\frac{1}{N}\sum_{n=1}^N (\Psi_\delta(x_n) - f(x_n))\Big|.\end{aligned}$$

Because of (4.9), the first term and the third term on the right-hand side are both $\leq \delta$ for all N. In view of (4.8), the second term on the right-hand side is $\leq \delta$ for sufficiently large N. Therefore

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^N f(\{x_n\}) = \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^N f(x_n) = \int_0^1 f(u)du.$$

Then the proof of Theorem 4.1.7 shows that the sequence $\{x_1\}, \{x_2\}, \dots$ of fractional parts is uniformly distributed in $[0, 1)$, and so the sequence $x_1, x_2, \dots$ is uniformly distributed modulo 1. If you don't like this proof via the Weierstrass approximation theorem, for instance because you have not seen this theorem before, then you can read the alternative proof in the last part of the proof of Theorem 4.1.14. □

The Weyl criterion affords an elegant way of proving that the sequence of multiples of an irrational number is uniformly distributed modulo 1. In fact, the following simple characterization can be established.

Theorem 4.1.10 *Let α be a real number. Then the sequence $\alpha, 2\alpha, \dots$ of multiples of α is uniformly distributed modulo 1 if and only if α is irrational.*

Proof. Let α be rational, say $\alpha = a/b$ with $a, b \in \mathbb{Z}$ and $b \geq 1$. Then none of the fractional parts $x_n = \{n\alpha\}$ with $n \in \mathbb{N}$ can be in the open interval $J = (0, 1/b)$, and so (4.7) is not satisfied for J. Therefore the sequence $x_1, x_2, \dots$ is not uniformly distributed in $[0,1)$, and so the sequence $\alpha, 2\alpha, \dots$ is not uniformly distributed modulo 1.

Now let α be irrational and let $h \in \mathbb{N}$. Then $e^{2\pi \mathrm{i} h\alpha} \neq 1$ and by the summation formula for geometric sums we get

$$\Big|\sum_{n=1}^{N} e^{2\pi \mathrm{i} hn\alpha}\Big| = \Big|\sum_{n=0}^{N-1} \left(e^{2\pi \mathrm{i} h\alpha}\right)^n\Big| = \frac{\left|e^{2\pi \mathrm{i} hN\alpha} - 1\right|}{\left|e^{2\pi \mathrm{i} h\alpha} - 1\right|} \leq \frac{2}{\left|e^{2\pi \mathrm{i} h\alpha} - 1\right|}.$$

It follows that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} e^{2\pi \mathrm{i} hn\alpha} = 0.$$

Therefore the sequence $\alpha, 2\alpha, \dots$ is uniformly distributed modulo 1 by the Weyl criterion. □

The sequence $\{\alpha\}, \{2\alpha\}, \dots$ of fractional parts of the multiples of an irrational number α is called a *Kronecker sequence* and was historically the first example of a uniformly distributed sequence. Kronecker sequences are named after Leopold Kronecker (1823–1891) who proved the first nontrivial result on them, namely that every Kronecker sequence is dense in the interval $[0,1]$ (that is, the points of a given Kronecker sequence come arbitrarily close to any point of $[0,1]$). The property of a Kronecker sequence of being uniformly distributed is of course stronger than that of being dense in $[0,1]$. Kronecker is famous also for the saying: "God made the integers, all else is the work of man."

Now that we know examples of uniformly distributed sequences, we can employ them in (4.6) to obtain convergent numerical integration schemes. In fact, we will see many more examples of uniformly distributed sequences in this chapter. The question of error bounds for the numerical integration scheme (4.6) leads to the concept of discrepancy which we introduce below.

As a matter of fact, there are two common notions of discrepancy, the star discrepancy $D_N^*(\mathcal{P})$ and the (extreme) discrepancy $D_N(\mathcal{P})$ of a point set $\mathcal{P}$ consisting of N points in $[0,1)$, and an easy connection exists between the two (see Proposition 4.1.12 below). The terminology *point set* designates what you would expect, namely a (finite) set of points, but there is the additional provision that the points can occur with a certain (finite) multiplicity. For instance, in the point set consisting of the four points $0, \frac{1}{2}, \frac{1}{2}, \frac{3}{4}$, the points 0 and $\frac{3}{4}$ occur with multiplicity 1 and the point $\frac{1}{2}$ occurs with multiplicity 2. The corresponding set (as opposed to point set) is $\{0, \frac{1}{2}, \frac{3}{4}\}$. The order in which the points of a point set are listed is irrelevant.

It is convenient to introduce a simple notation for the sum occurring in (4.7); namely, for a point set $\mathcal{P}$ consisting of $x_1, \dots, x_N \in [0,1)$ and for a subinterval J of $[0,1]$, we write

$$A(J;\mathcal{P}) = \sum_{n=1}^{N} c_J(x_n).$$

In words, $A(J;\mathcal{P})$ is the number of integers n with $1 \le n \le N$ such that $x_n \in J$. Note that the multiplicities of points in $\mathcal{P}$ are taken into account when computing the counting function $A(J;\mathcal{P})$. For instance, for the point set $\mathcal{P}$ consisting of $0, \frac{1}{2}, \frac{1}{2}, \frac{3}{4}$ as above and for $J = [0, \frac{3}{4})$, the point 0 with multiplicity 1 and the point $\frac{1}{2}$ with multiplicity 2 lie in J, and so $A(J;\mathcal{P}) = 1 + 2 = 3$.

Definition 4.1.11 Let $\mathcal{P}$ be the point set consisting of the N points $x_1, \ldots, x_N \in [0,1)$. Then the *star discrepancy* $D_N^*(\mathcal{P})$ of $\mathcal{P}$ is defined by

$$D_N^*(\mathcal{P}) = \sup_{0<u\le 1} \Big| \frac{A([0,u);\mathcal{P})}{N} - u \Big|$$

and the (*extreme*) *discrepancy* $D_N(\mathcal{P})$ of $\mathcal{P}$ is defined by

$$D_N(\mathcal{P}) = \sup_{0\le u<v\le 1} \Big| \frac{A([u,v);\mathcal{P})}{N} - (v-u) \Big|.$$

The idea of the star discrepancy and of the (extreme) discrepancy can be comprehended in terms of the limit relation (4.7), since for finite N these discrepancies tell us how close we are to the limit on the right-hand side of (4.7) in the worst case. It is trivial that always $D_N^*(\mathcal{P}) \le 1$ and $D_N(\mathcal{P}) \le 1$.

Proposition 4.1.12 *Every point set $\mathcal{P}$ consisting of N points in $[0,1)$ satisfies*

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

Proof. The first inequality follows immediately from the definitions. Next we note that $A([u,v);\mathcal{P}) = A([0,v);\mathcal{P}) - A([0,u);\mathcal{P})$, where $0 \le u < v \le 1$, and therefore

$$\Big| \frac{A([u,v);\mathcal{P})}{N} - (v-u) \Big| \le \Big| \frac{A([0,v);\mathcal{P})}{N} - v \Big| + \Big| \frac{A([0,u);\mathcal{P})}{N} - u \Big|.$$

Taking suprema yields the second inequality in the proposition. □

In practice one is often interested in the order of magnitude of $D_N^*(\mathcal{P})$ and $D_N(\mathcal{P})$ as a function of N. Proposition 4.1.12 shows that from this perspective it does not matter which of the two discrepancies we consider.

Probably the most important result on the discrepancy in terms of the number of applications it allows is the following bound on the discrepancy by means of exponential sums. This bound can be viewed as a quantitative version of the Weyl criterion in Theorem 4.1.9.

Theorem 4.1.13 (Erdős-Turán Inequality) *If $\mathcal{P}$ is the point set consisting of the N points $x_1, \ldots, x_N \in [0,1)$, then*

$$D_N(\mathcal{P}) \le \frac{6}{H+1} + \frac{4}{\pi} \sum_{h=1}^{H} \Big(\frac{1}{h} - \frac{1}{H+1} \Big) \Big| \frac{1}{N} \sum_{n=1}^{N} e^{2\pi i h x_n} \Big|$$

for all positive integers H.

Proof. We introduce the function

$$R(u) = \frac{A([0,u);\mathcal{P})}{N} - u = \frac{1}{N}\sum_{n=1}^{N} c_{[0,u)}(x_n) - u \qquad \text{for } 0 \le u \le 1.$$

Since $R(0) = 0 = R(1)$, we can extend this function to $\mathbb{R}$ with period 1. Next we put

$$r(u) = R(u) - \int_0^1 R(u)du \qquad \text{for all } u \in \mathbb{R}$$

and we note that

$$\int_0^1 r(u)du = 0. \tag{4.10}$$

A straightforward computation shows that for every nonzero integer h,

$$\begin{aligned}\int_0^1 r(u)e^{2\pi i hu}du &= \frac{1}{N}\sum_{n=1}^{N}\int_0^1 c_{[0,u)}(x_n)e^{2\pi i hu}du - \int_0^1 ue^{2\pi i hu}du \\ &= \frac{1}{N}\sum_{n=1}^{N}\int_{x_n}^1 e^{2\pi i hu}du - \frac{1}{2\pi i h} \\ &= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{2\pi i h}(1 - e^{2\pi i h x_n}) - \frac{1}{2\pi i h} = -\frac{S_h}{2\pi i h}\end{aligned}$$

with $S_h = (1/N)\sum_{n=1}^{N} e^{2\pi i h x_n}$.

Fix a positive integer H and let a be a real number to be determined later. By forming an appropriate linear combination of the last displayed identity and using (4.10), we obtain

$$\begin{aligned}-\sum_{h=-H}^{H}{}^{*}(H+1-|h|)e^{-2\pi i ha}\frac{S_h}{2\pi i h} &= \int_0^1 r(u)\Big(\sum_{h=-H}^{H}(H+1-|h|)e^{2\pi i h(u-a)}\Big)du \\ &= \int_{-a}^{1-a} r(u+a)\Big(\sum_{h=-H}^{H}(H+1-|h|)e^{2\pi i hu}\Big)du,\end{aligned}$$

where the asterisk indicates that $h = 0$ is deleted from the range of summation. Because of the periodicity of the integrand, the last integral may also be taken over the interval $[-\frac{1}{2}, \frac{1}{2}]$, and so we can write

$$\int_{-1/2}^{1/2} r(u+a)\Big(\sum_{h=-H}^{H}(H+1-|h|)e^{2\pi i hu}\Big)du = -\sum_{h=-H}^{H}{}^{*}(H+1-|h|)\frac{e^{-2\pi i ha}S_h}{2\pi i h}. \tag{4.11}$$

Elementary trigonometry shows that

$$e^{-\pi i Hu}\sum_{h=0}^{H} e^{2\pi i hu} = \frac{\sin(H+1)\pi u}{\sin \pi u},$$

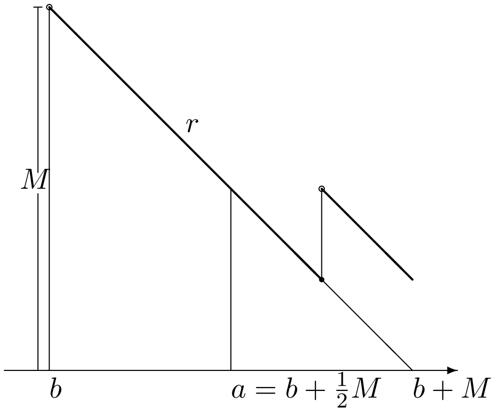

Figure 4.3: The graph of r

where the right-hand side is interpreted as $H+1$ if $u \in \mathbb{Z}$. By squaring this identity, we obtain

$$\frac{\sin^2(H+1)\pi u}{\sin^2 \pi u} = e^{-2\pi \mathrm{i} H u}\Big(\sum_{h=0}^{H} e^{2\pi \mathrm{i} h u}\Big)^2 = \sum_{h=-H}^{H} (H+1-|h|)e^{2\pi \mathrm{i} h u}. \tag{4.12}$$

Now we insert this formula into (4.11) and apply the triangle inequality to get

$$\left|\int_{-1/2}^{1/2} r(u+a)\frac{\sin^2(H+1)\pi u}{\sin^2 \pi u} du\right| \le \frac{1}{2\pi}\sum_{h=-H}^{H}{}^{*}\Big(\frac{H+1}{|h|}-1\Big)|S_h|.$$

We observe that $|S_{-h}| = |S_h|$ for every $h \in \mathbb{N}$, and so

$$\left|\int_{-1/2}^{1/2} r(u+a)\frac{\sin^2(H+1)\pi u}{\sin^2 \pi u} du\right| \le \frac{1}{\pi}\sum_{h=1}^{H}\Big(\frac{H+1}{h}-1\Big)|S_h|. \tag{4.13}$$

Next we put

$$M = \sup_{u \in \mathbb{R}} |r(u)|.$$

The function r is continuous from the left, has only positive jumps, and is piecewise linear with slope -1. This implies that either $r(b) = -M$ or $r(b+0) = M$ for some $b \in \mathbb{R}$. We treat only the second alternative, the first one being completely similar. For $b < t \le b + M$, the properties of the function r yield

$$r(t) = M + r(t) - r(b+0) \ge M + b - t.$$

Now we choose $a = b + \frac{1}{2}M$ (see Figure 4.3). Then the inequality above with $t = b + \frac{1}{2}M + u$ implies that

$$r(u+a) \ge \frac{1}{2}M - u \qquad \text{for } |u| < \frac{1}{2}M.$$

This shows in particular that $M \leq 1$. Consequently, we get

$$\begin{aligned}
&\int_{-1/2}^{1/2} r(u+a)\frac{\sin^2(H+1)\pi u}{\sin^2 \pi u}du \\
&= \left(\int_{-M/2}^{M/2} + \int_{-1/2}^{-M/2} + \int_{M/2}^{1/2}\right) r(u+a)\frac{\sin^2(H+1)\pi u}{\sin^2 \pi u}du \\
&\geq \int_{-M/2}^{M/2} \left(\frac{1}{2}M - u\right) \frac{\sin^2(H+1)\pi u}{\sin^2 \pi u}du \\
&\quad -M\int_{-1/2}^{-M/2} \frac{\sin^2(H+1)\pi u}{\sin^2 \pi u}du - M\int_{M/2}^{1/2} \frac{\sin^2(H+1)\pi u}{\sin^2 \pi u}du.
\end{aligned}$$

Now we use the evenness of the function $(\sin^2(H+1)\pi u)/(\sin^2 \pi u)$ to obtain

$$\begin{aligned}
&\int_{-1/2}^{1/2} r(u+a)\frac{\sin^2(H+1)\pi u}{\sin^2 \pi u}du \\
&\geq M\int_{0}^{M/2} \frac{\sin^2(H+1)\pi u}{\sin^2 \pi u}du - 2M\int_{M/2}^{1/2} \frac{\sin^2(H+1)\pi u}{\sin^2 \pi u}du \\
&= M\int_{0}^{1/2} \frac{\sin^2(H+1)\pi u}{\sin^2 \pi u}du - 3M\int_{M/2}^{1/2} \frac{\sin^2(H+1)\pi u}{\sin^2 \pi u}du.
\end{aligned}$$

By applying again the evenness of the function $(\sin^2(H+1)\pi u)/(\sin^2 \pi u)$ as well as (4.12), we get

$$\int_{0}^{1/2} \frac{\sin^2(H+1)\pi u}{\sin^2 \pi u}du = \frac{1}{2}\int_{-1/2}^{1/2} \frac{\sin^2(H+1)\pi u}{\sin^2 \pi u}du = \frac{H+1}{2},$$

and so

$$\int_{-1/2}^{1/2} r(u+a)\frac{\sin^2(H+1)\pi u}{\sin^2 \pi u}du \geq \frac{H+1}{2}M - 3M\int_{M/2}^{1/2} \frac{du}{\sin^2 \pi u}.$$

Now $\sin \pi u \geq 2u$ for $0 \leq u \leq \frac{1}{2}$, and this yields

$$\int_{M/2}^{1/2} \frac{du}{\sin^2 \pi u} \leq \int_{M/2}^{1/2} \frac{du}{4u^2} = \frac{1}{2M} - \frac{1}{2}.$$

Therefore

$$\int_{-1/2}^{1/2} r(u+a)\frac{\sin^2(H+1)\pi u}{\sin^2 \pi u}du \geq \frac{H+1}{2}M - \frac{3}{2}.$$

By combining this with (4.13), we arrive at the bound

$$M \leq \frac{3}{H+1} + \frac{2}{\pi}\sum_{h=1}^{H}\Big(\frac{1}{h} - \frac{1}{H+1}\Big)|S_h|.$$

We note that

$$D_N(\mathcal{P}) = \sup_{u,v\in\mathbb{R}} |R(v) - R(u)| = \sup_{u,v\in\mathbb{R}} |r(v) - r(u)| \leq 2M,$$

and this proves the Erdős-Turán inequality. □

For a sequence $\mathcal{S}$ of real numbers $x_1, x_2, \ldots$, we will now often write $\mathcal{S} = (x_n)_{n=1}^{\infty}$. If the x_n are in $[0,1)$, then for every positive integer N let $D_N^*(\mathcal{S})$, respectively $D_N(\mathcal{S})$, be the star discrepancy, respectively discrepancy, of the first N terms $x_1, \ldots, x_N$ of $\mathcal{S}$. There is a simple relationship between the uniform distribution of $\mathcal{S}$ and the asymptotic behavior of these discrepancies.

Theorem 4.1.14 *The following properties of a sequence $\mathcal{S}$ of points in $[0,1)$ are equivalent:*
(i) *$\mathcal{S}$ is uniformly distributed in $[0,1)$;*
(ii) $\lim_{N\to\infty} D_N^*(\mathcal{S}) = 0$*;*
(iii) $\lim_{N\to\infty} D_N(\mathcal{S}) = 0$.

Proof. The properties (ii) and (iii) are equivalent since $D_N^*(\mathcal{S}) \leq D_N(\mathcal{S}) \leq 2D_N^*(\mathcal{S})$ for every $N \in \mathbb{N}$ by Proposition 4.1.12. Thus, it remains to show that (i) and (iii) are equivalent. The fact that (iii) implies (i) is trivial in view of Theorem 4.1.6. Finally, suppose that (i) is satisfied, so that $\mathcal{S} = (x_n)_{n=1}^{\infty}$ is uniformly distributed in $[0,1)$. Then Theorem 4.1.9 yields

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} e^{2\pi \mathrm{i} h x_n} = 0 \qquad \text{for all } h \in \mathbb{N}.$$

Now we fix an integer $H \geq 1$ and we let $N \to \infty$ in the Erdős-Turán inequality (see Theorem 4.1.13). This yields

$$0 \leq \liminf_{N\to\infty} D_N(\mathcal{S}) \leq \limsup_{N\to\infty} D_N(\mathcal{S}) \leq \frac{6}{H+1}.$$

Since H can be arbitrarily large, we infer that $\lim_{N\to\infty} D_N(\mathcal{S}) = 0$. □

We continue with some simple observations about the star discrepancy which is the main tool for obtaining error bounds in (4.2).

Lemma 4.1.15 *Let $\mathcal{P}_1$ be the point set consisting of $x_1, \ldots, x_N \in [0,1)$ and let $\mathcal{P}_2$ be the point set consisting of $y_1, \ldots, y_N \in [0,1)$. Suppose that for some $\varepsilon > 0$ the inequality $|x_n - y_n| \leq \varepsilon$ holds for $1 \leq n \leq N$. Then*

$$|D_N^*(\mathcal{P}_1) - D_N^*(\mathcal{P}_2)| \leq \varepsilon.$$

Proof. Consider any interval $J = [0, u) \subseteq [0, 1)$. Whenever $y_n \in J$, then $x_n \in J_1 := [0, \min(u + \varepsilon, 1))$; hence

$$\frac{A(J; \mathcal{P}_2)}{N} - \lambda(J) \leq \frac{A(J_1; \mathcal{P}_1)}{N} - \lambda(J_1) + \varepsilon \leq D_N^*(\mathcal{P}_1) + \varepsilon.$$

Whenever $x_n \in J_2 := [0, \max(u - \varepsilon, 0))$, then $y_n \in J$; hence

$$\frac{A(J; \mathcal{P}_2)}{N} - \lambda(J) \geq \frac{A(J_2; \mathcal{P}_1)}{N} - \lambda(J_2) - \varepsilon \geq -D_N^*(\mathcal{P}_1) - \varepsilon.$$

Thus $D_N^*(\mathcal{P}_2) \leq D_N^*(\mathcal{P}_1) + \varepsilon$. By interchanging the roles of $\mathcal{P}_1$ and $\mathcal{P}_2$, we obtain $D_N^*(\mathcal{P}_1) \leq D_N^*(\mathcal{P}_2) + \varepsilon$, and so $|D_N^*(\mathcal{P}_1) - D_N^*(\mathcal{P}_2)| \leq \varepsilon$. □

The following is a nice explicit formula for the star discrepancy. Since the star discrepancy of a point set does not depend on the order in which the points of the point set are listed, we can arrange them in nondecreasing order.

Proposition 4.1.16 *Let $\mathcal{P}$ be the point set consisting of $x_1, \ldots, x_N \in [0, 1)$ and suppose that $x_1 \leq x_2 \leq \cdots \leq x_N$. Then*

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

Proof. Since $D_N^*(\mathcal{P})$ is a continuous function of $x_1, \ldots, x_N$ by Lemma 4.1.15, we can assume that $0 < x_1 < x_2 < \cdots < x_N < 1$. Put $x_0 = 0$ and $x_{N+1} = 1$. Then simple considerations show that

$$\begin{aligned} D_N^*(\mathcal{P}) &= \max_{0 \leq n \leq N} \sup_{x_n < u \leq x_{n+1}} \left| \frac{A([0, u); \mathcal{P})}{N} - u \right| \\ &= \max_{0 \leq n \leq N} \sup_{x_n < u \leq x_{n+1}} \left| \frac{n}{N} - u \right| \\ &= \max_{0 \leq n \leq N} \max \left(\left| \frac{n}{N} - x_n \right|, \left| \frac{n}{N} - x_{n+1} \right| \right) \\ &= \max_{1 \leq n \leq N} \max \left(\left| \frac{n}{N} - x_n \right|, \left| \frac{n-1}{N} - x_n \right| \right). \end{aligned}$$

Now

$$\max \left(\left| \frac{n}{N} - x_n \right|, \left| \frac{n-1}{N} - x_n \right| \right) = \frac{1}{2N} + \left| x_n - \frac{2n-1}{2N} \right| \qquad \text{for } 1 \leq n \leq N,$$

and this yields the desired formula for $D_N^*(\mathcal{P})$. □

Corollary 4.1.17 *Every point set $\mathcal{P}$ consisting of N points in $[0, 1)$ satisfies*

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

Proof. This follows immediately from Proposition 4.1.16. □

Remark 4.1.18 The formula for $D_N^*(\mathcal{P})$ in Proposition 4.1.16 implies also that $D_N^*(\mathcal{P}) = \frac{1}{2N}$ if and only if the points $x_1, \ldots, x_N$ form a permutation of the points $\frac{1}{2N}, \frac{3}{2N}, \ldots, \frac{2N-1}{2N}$. It is of interest to observe that the latter points are exactly the nodes of the midpoint rule with N nodes.

By using a different approach, we can prove a lower bound on $D_N(\mathcal{P})$ as well, and thus we get the following companion result to Corollary 4.1.17.

Proposition 4.1.19 *Every point set $\mathcal{P}$ consisting of N points in $[0,1)$ satisfies*

$$D_N(\mathcal{P}) \geq \frac{1}{N}.$$

Proof. Let $x \in [0,1)$ be any point of $\mathcal{P}$. We choose $\varepsilon > 0$ and consider the half-open interval $J = [x-\varepsilon, x+\varepsilon) \cap [0,1)$. Since $x \in J$, we get

$$\frac{A(J;\mathcal{P})}{N} - \lambda(J) \geq \frac{1}{N} - \lambda(J) \geq \frac{1}{N} - 2\varepsilon,$$

and so $D_N(\mathcal{P}) \geq \frac{1}{N} - 2\varepsilon$. The desired bound is obtained by letting $\varepsilon \to 0+$. □

Example 4.1.20 For the point set $\mathcal{P}$ in Remark 4.1.18 consisting of $\frac{1}{2N}, \frac{3}{2N}, \ldots, \frac{2N-1}{2N}$, we obtain $D_N(\mathcal{P}) \leq 2D_N^*(\mathcal{P}) = \frac{1}{N}$, and so $D_N(\mathcal{P}) = \frac{1}{N}$ by Proposition 4.1.19. There are further examples of N points in $[0,1)$ with discrepancy $\frac{1}{N}$, for instance, the equidistant points $0, \frac{1}{N}, \ldots, \frac{N-1}{N}$.

We now return to the general form of an equal-weight rule for $[0,1]$ which, according to (4.2), provides the approximation

$$\int_0^1 f(u)du \approx \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

with nodes $x_1, \ldots, x_N \in [0,1]$. This numerical integration scheme is also called *quasi-Monte Carlo integration* since it is a simple instance of a *quasi-Monte Carlo method*, that is, a deterministic version of a Monte Carlo method (see Subsection 4.1.2). Error bounds for quasi-Monte Carlo integration can be given in terms of the star discrepancy of the nodes. Historically the first such error bound is the following inequality of Koksma [84] for integrands of bounded variation. Recall that for a real-valued function f on $[0,1]$, its *variation* $V(f)$ is defined to be

$$V(f) = \sup \sum_{i=0}^{m-1} |f(y_{i+1}) - f(y_i)|,$$

where the supremum is extended over all real numbers $0 = y_0 < y_1 < \cdots < y_m = 1$ with an arbitrary $m \in \mathbb{N}$, and f has *bounded variation* if $V(f) < \infty$.

Theorem 4.1.21 (Koksma Inequality) *If the real-valued function f has bounded variation $V(f)$ on $[0,1]$ and $x_1, \dots, x_N \in [0,1)$ are arbitrary, then*

$$\left| \int_0^1 f(u)du - \frac{1}{N} \sum_{n=1}^{N} f(x_n) \right| \le V(f) D_N^*(\mathcal{P}),$$

where $D_N^*(\mathcal{P})$ is the star discrepancy of the point set $\mathcal{P}$ consisting of $x_1, \dots, x_N$.

Proof. We can assume that $x_1 \le x_2 \le \dots \le x_N$. Put $x_0 = 0$ and $x_{N+1} = 1$. Using integration by parts and summation by parts, we obtain

$$\begin{aligned} \int_0^1 f(u)du - \frac{1}{N} \sum_{n=1}^{N} f(x_n) &= - \int_0^1 u df(u) + \sum_{n=0}^{N} \frac{n}{N} (f(x_{n+1}) - f(x_n)) \\ &= \sum_{n=0}^{N} \int_{x_n}^{x_{n+1}} \Big(\frac{n}{N} - u \Big) df(u). \end{aligned}$$

For fixed n with $0 \le n \le N$, we get

$$\left| \frac{n}{N} - u \right| \le \max \left(\left| x_n - \frac{n}{N} \right|, \left| x_{n+1} - \frac{n}{N} \right| \right) \le D_N^*(\mathcal{P}) \qquad \text{for } x_n \le u \le x_{n+1}$$

by Proposition 4.1.16, and so the desired inequality follows. □

Remark 4.1.22 If f has a continuous first derivative f' on $[0,1]$, then we can use $df(u) = f'(u)du$ in the proof of Theorem 4.1.21. Then the proof can be written in terms of ordinary Riemann integrals, namely

$$\begin{aligned} \left| \int_0^1 f(u)du - \frac{1}{N} \sum_{n=1}^{N} f(x_n) \right| &\le \sum_{n=0}^{N} \left| \int_{x_n}^{x_{n+1}} \Big(\frac{n}{N} - u \Big) f'(u)du \right| \\ &\le \sum_{n=0}^{N} D_N^*(\mathcal{P}) \int_{x_n}^{x_{n+1}} |f'(u)|du = D_N^*(\mathcal{P}) \int_0^1 |f'(u)|du. \end{aligned}$$

Now $\int_0^1 |f'(u)|du = V(f)$ under the given condition on f, and so we obtain the same bound as in Theorem 4.1.21.

Remark 4.1.23 If $\mathcal{S} = (x_n)_{n=1}^{\infty}$ is a uniformly distributed sequence in $[0,1)$ and $\mathcal{P}_N$ is the point set consisting of the first N terms $x_1, \dots, x_N$ of $\mathcal{S}$, then $D_N^*(\mathcal{P}_N) = D_N^*(\mathcal{S}) \to 0$ as $N \to \infty$ by Theorem 4.1.14. Hence the error bound $V(f)D_N^*(\mathcal{P}_N)$ in Theorem 4.1.21 also tends to 0 as $N \to \infty$.

Continuous functions need not be of bounded variation; for instance, the function f with $f(u) = u \sin(1/u)$ for $0 < u \le 1$ and $f(0) = 0$ is continuous on $[0,1]$, but not of bounded variation on $[0,1]$. Therefore it is of interest to establish also an error bound for quasi-Monte Carlo integration with continuous integrands. Such an error bound was shown in [124] and it uses the modulus of continuity of the integrand. For a real-valued continuous function f on $[0,1]$, its *modulus of continuity* is defined by

$$M(f;t) = \sup_{\substack{u,v\in[0,1] \\ |u-v|\le t}} |f(u)-f(v)| \qquad \text{for all } t \ge 0.$$

Theorem 4.1.24 *If f is a real-valued continuous function on $[0,1]$ and $x_1,\ldots,x_N \in [0,1)$ are arbitrary, then*

$$\left|\int_0^1 f(u)du - \frac{1}{N}\sum_{n=1}^{N} f(x_n)\right| \le M(f;D_N^*(\mathcal{P})),$$

where $D_N^*(\mathcal{P})$ is the star discrepancy of the point set $\mathcal{P}$ consisting of the points $x_1,\ldots,x_N$.

Proof. We can assume that $x_1 \le x_2 \le \cdots \le x_N$. The mean-value theorem for integrals allows us to write

$$\int_0^1 f(u)du = \sum_{n=1}^{N}\int_{(n-1)/N}^{n/N} f(u)du = \frac{1}{N}\sum_{n=1}^{N} f(\xi_n)$$

with $(n-1)/N \le \xi_n \le n/N$ for $1 \le n \le N$. Therefore

$$\int_0^1 f(u)du - \frac{1}{N}\sum_{n=1}^{N} f(x_n) = \frac{1}{N}\sum_{n=1}^{N}(f(\xi_n)-f(x_n)).$$

For every n with $1 \le n \le N$, we obtain

$$|\xi_n - x_n| \le \max\left(\left|x_n - \frac{n-1}{N}\right|, \left|x_n - \frac{n}{N}\right|\right) \le D_N^*(\mathcal{P})$$

by Proposition 4.1.16, and so

$$\left|\int_0^1 f(u)du - \frac{1}{N}\sum_{n=1}^{N} f(x_n)\right| \le \frac{1}{N}\sum_{n=1}^{N}|f(\xi_n)-f(x_n)| \le M(f;D_N^*(\mathcal{P}))$$

as desired. □

Remark 4.1.25 Every real-valued continuous function f on the compact interval $[0,1]$ is uniformly continuous, and so its modulus of continuity satisfies $M(f;t) \to 0$ as $t \to 0+$. Therefore in the situation considered in Remark 4.1.23, the error bound $M(f;D_N^*(\mathcal{P}_N))$ in Theorem 4.1.24 tends to 0 as $N \to \infty$.

4.1.2 The multidimensional case

Numerical integration in the one-dimensional case is considered essentially a solved problem since classical numerical integration rules do a good job for most of the one-dimensional integrals arising in practice. The greater challenge in numerical integration is the multidimensional case, particularly if the dimension is high. There are important practical applications where the dimension of the integral to be computed goes into the hundreds or even thousands, with computational finance perhaps being the area that produces the greatest number of high-dimensional numerical integration problems. The main task of computational finance is the calculation of the monetary values of sophisticated financial instruments such as stock options. A coverage of computational finance is beyond the scope of this book; we refer instead to the comprehensive treatise by Glasserman [55].

We standardize the multidimensional numerical integration problem by considering, for a given dimension $s \geq 2$, a definite integral $\int_{[0,1]^s} f(\mathbf{u})d\mathbf{u}$ over the s-dimensional unit cube $[0,1]^s$ with integration variable $\mathbf{u} = (u_1, \ldots, u_s)$. The classical approach to this multidimensional numerical integration problem uses Cartesian products of one-dimensional integration rules. In such multidimensional integration rules, the node set is a Cartesian product of one-dimensional node sets and the weights are products of corresponding weights taken from the one-dimensional rules. These multidimensional integration rules are obtained by viewing the given s-dimensional integral

$$\int_{[0,1]^s} f(\mathbf{u})d\mathbf{u} = \int_0^1 \cdots \int_0^1 f(u_1, \ldots, u_s)du_1 \cdots du_s$$

as an iteration of one-dimensional integrals and by applying a one-dimensional integration rule in each iteration.

We illustrate this procedure with the s-fold Cartesian product of the midpoint rule (4.3). If we apply the midpoint rule with $m \geq 1$ nodes, then the s-fold Cartesian product attains the form

$$\int_{[0,1]^s} f(\mathbf{u})d\mathbf{u} \approx \frac{1}{m^s} \sum_{k_1=1}^{m} \cdots \sum_{k_s=1}^{m} f\Big(\frac{2k_1 - 1}{2m}, \ldots, \frac{2k_s - 1}{2m}\Big). \tag{4.14}$$

The total number of nodes in (4.14) is $N = m^s$. From the error bound for the midpoint rule in (4.4) it follows easily that the error in (4.14) is $O(m^{-2})$, provided that the partial derivatives $\partial^2 f / \partial u_i^2$ are continuous on $[0,1]^s$ for $1 \leq i \leq s$. In order to see that the error in (4.14) need not, in general, be smaller than the one-dimensional integration error, it suffices to apply (4.14) with a function f on $[0,1]^s$ that depends on only one variable, in which case (4.14) reduces to (4.3).

In terms of the total number $N = m^s$ of nodes in (4.14), the error bound $O(m^{-2})$ in (4.14) is in fact $O(N^{-2/s})$. With increasing dimension s, the utility of the error bound $O(N^{-2/s})$ declines drastically. Specifically, in order to guarantee a prescribed level of accuracy, say an error that is in absolute value at most 10^{-2}, we must use roughly 10^s nodes. Therefore the required number of nodes increases exponentially with the dimension s, so that even for moderately large s the computation may become infeasible. This phenomenon is often called the *curse of dimensionality*. The curse of dimensionality manifests itself in an analogous way for the Cartesian product of any one-dimensional integration rule. For an s-fold Cartesian product, the order of magnitude of the error bound, in terms of the total number of nodes, is the sth root of the order of magnitude of the error bound for the one-dimensional integration rule.

A technique to overcome the curse of dimensionality is the *Monte Carlo method*. Just to be sure, this is *not* a foolproof scheme to win at roulette in Monte Carlo, but a numerical method based on random sampling. The Monte Carlo method has a fascinating history which goes back at least to the 1940s and involves famous mathematicians like John von Neumann and Stanislaw Ulam; see [60, Section 1.2] for a brief history of the Monte Carlo method and [42] for an account of the work of von Neumann and Ulam on the Monte Carlo method at the Los Alamos Scientific Laboratory. Since the Monte Carlo method was developed in the United States, it could just as well have been called the Las Vegas method, but several of the co-inventors of the method were of European origin (for instance, von Neumann came from Hungary and Ulam from Poland) and they preferred Monte Carlo to Las Vegas.

The Monte Carlo method is a widely applicable computational tool, but we consider it only in the context of numerical integration. Informative textbooks on the general Monte Carlo method are Fishman [51] and Kalos and Whitlock [76] (and also [119] if you know German), while Glasserman [55], Lemieux [97], and Leobacher and Pillichshammer [98] discuss applications to computational finance.

In the Monte Carlo method for the numerical integration of our standard s-dimensional integral $\int_{[0,1]^s} f(\mathbf{u})d\mathbf{u}$, we use the *Monte Carlo estimate*

$$\int_{[0,1]^s} f(\mathbf{u})d\mathbf{u} \approx \frac{1}{N}\sum_{n=1}^{N} f(\mathbf{x}_n), \tag{4.15}$$

where $\mathbf{x}_1, \dots, \mathbf{x}_N$ are independent and uniformly distributed random samples from $[0,1]^s$ (in the sense of statistics). In the language of statistics, if f is Riemann-integrable (or even only Lebesgue-integrable) on $[0,1]^s$, then the left-hand side of (4.15) is the expected value of f as a random variable and the right-hand side of (4.15) is the sample average. Approximating an expected value by a sample average as in (4.15) is a plausible principle in statistics, supported by what statisticians call the law of large numbers. This law was poetically expressed by the writer Tom Stoppard in his play *Rosencrantz and Guildenstern Are Dead*: "[This law] related the fortuitous and the ordained into a reassuring union which we recognized as nature. The sun came up about as often as it went down, in the long run."

It should be evident that the error analysis for the Monte Carlo estimate will, due to its statistical nature, proceed by statistical and probabilistic arguments. Since statistics and probability theory are not prerequisites for this book, we state the results of the error analysis informally and without proof (see the textbooks on the Monte Carlo method mentioned above for rigorous statements and proofs). First of all, if f is (Lebesgue-)integrable, then with probability 1 we get the limit relation

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} f(\mathbf{x}_n) = \int_{[0,1]^s} f(\mathbf{u})d\mathbf{u},$$

that is, the Monte Carlo method converges. If not only f, but also f^2 is integrable, then with positive probability the error bound

$$\int_{[0,1]^s} f(\mathbf{u})d\mathbf{u} - \frac{1}{N}\sum_{n=1}^{N} f(\mathbf{x}_n) = O(N^{-1/2}) \tag{4.16}$$

is valid, and we can push the probability as close to 1 as we want if we choose the implied constant on the right-hand side of (4.16) sufficiently large. The remarkable feature of the order of magnitude $N^{-1/2}$ of the error bound in (4.16) is that it does not depend on the dimension s. Consequently, the Monte Carlo method allows us to beat the curse of dimensionality.

The number-theoretic approach to multidimensional numerical integration proceeds, as in the one-dimensional case described in Subsection 4.1.1, by the theory of uniform distribution of sequences. Let us start the ball rolling by generalizing Definition 4.1.3 to the multidimensional case.

Definition 4.1.26 A sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots$ of points in the half-open s-dimensional unit cube $[0,1)^s$ is *uniformly distributed* (in $[0,1)^s$) if

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} f(\mathbf{x}_n) = \int_{[0,1]^s} f(\mathbf{u})d\mathbf{u}$$

for every real-valued Riemann-integrable function f on $[0,1]^s$.

The criteria for uniform distribution in $[0,1)$ established in Subsection 4.1.1 can be extended to the multidimensional case in a straightforward manner (see [41, Subsection 1.1.1] and [91, Section 1.6] for the details). For a subinterval J of $[0,1]^s$, we write c_J for the characteristic function of J, that is, $c_J(\mathbf{u}) = 1$ if $\mathbf{u} \in J$ and $c_J(\mathbf{u}) = 0$ if $\mathbf{u} \in [0,1]^s \setminus J$. Let $\lambda_s(J)$ denote the s-dimensional volume of J (if you are familiar with the Lebesgue measure, then you may also think of λ_s as the s-dimensional Lebesgue measure).

Theorem 4.1.27 *A sequence* $\mathbf{x}_1, \mathbf{x}_2, \ldots$ *of points in* $[0,1)^s$ *is uniformly distributed in* $[0,1)^s$ *if and only if*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} c_J(\mathbf{x}_n) = \lambda_s(J)$$

for every subinterval J *of* $[0,1]^s$*.*

Theorem 4.1.28 *A sequence* $\mathbf{x}_1, \mathbf{x}_2, \ldots$ *of points in* $[0,1)^s$ *is uniformly distributed in* $[0,1)^s$ *if and only if*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(\mathbf{x}_n) = \int_{[0,1]^s} f(\mathbf{u}) d\mathbf{u}$$

for every real-valued continuous function f *on* $[0,1]^s$*.*

For a point $\mathbf{x} = (x_1, \ldots, x_s) \in \mathbb{R}^s$, the *fractional part* $\{\mathbf{x}\}$ is defined by

$$\{\mathbf{x}\} = (\{x_1\}, \ldots, \{x_s\}) \in [0,1)^s,$$

where $\{x\} = x - \lfloor x \rfloor$ denotes as in Subsection 4.1.1 the fractional part of the real number x.

Definition 4.1.29 A sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots$ of points in $\mathbb{R}^s$ is *uniformly distributed modulo* 1 *in* $\mathbb{R}^s$ if the sequence $\{\mathbf{x}_1\}, \{\mathbf{x}_2\}, \ldots$ of fractional parts is uniformly distributed in $[0,1)^s$.

The Weyl criterion in Theorem 4.1.9 can also be generalized to the multidimensional case (see [41, Theorem 1.19] and [91, Section 1.6] for the details). For points $\mathbf{x} = (x_1, \ldots, x_s)$ and $\mathbf{y} = (y_1, \ldots, y_s)$ in $\mathbb{R}^s$, we write

$$\mathbf{x} \cdot \mathbf{y} = x_1 y_1 + \cdots + x_s y_s$$

for the dot product (or standard inner product) on $\mathbb{R}^s$. It is again convenient to abbreviate a sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots$ of points in $\mathbb{R}^s$ by $(\mathbf{x}_n)_{n=1}^{\infty}$.

Theorem 4.1.30 (Weyl Criterion in $\mathbb{R}^s$) *A sequence* $(\mathbf{x}_n)_{n=1}^{\infty}$ *of points in* $\mathbb{R}^s$ *is uniformly distributed modulo* 1 *in* $\mathbb{R}^s$ *if and only if*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} e^{2\pi \mathrm{i} \mathbf{h} \cdot \mathbf{x}_n} = 0$$

for every lattice point $\mathbf{h} \in \mathbb{Z}^s$ *with* $\mathbf{h} \neq \mathbf{0}$*.*

Corollary 4.1.31 *A sequence* $(\mathbf{x}_n)_{n=1}^{\infty}$ *of points in* $\mathbb{R}^s$ *is uniformly distributed modulo* 1 *in* $\mathbb{R}^s$ *if and only if, for every lattice point* $\mathbf{h} \in \mathbb{Z}^s$ *with* $\mathbf{h} \neq \mathbf{0}$*, the sequence* $(\mathbf{h} \cdot \mathbf{x}_n)_{n=1}^{\infty}$ *of dot products is uniformly distributed modulo* 1*.*

Proof. This follows immediately from Theorems 4.1.30 and 4.1.9. □

We use Corollary 4.1.31 to prove the following multidimensional version of Theorem 4.1.10.

Theorem 4.1.32 *For $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_s) \in \mathbb{R}^s$, the sequence $(n\boldsymbol{\alpha})_{n=1}^{\infty}$ of multiples of $\boldsymbol{\alpha}$ is uniformly distributed modulo 1 in $\mathbb{R}^s$ if and only if the real numbers $1, \alpha_1, \ldots, \alpha_s$ are linearly independent over the field $\mathbb{Q}$ of rational numbers.*

Proof. Put $\mathbf{x}_n = n\boldsymbol{\alpha}$ for $n = 1, 2, \ldots$. Suppose first that $1, \alpha_1, \ldots, \alpha_s$ are linearly independent over $\mathbb{Q}$. For every $\mathbf{h} \in \mathbb{Z}^s$ with $\mathbf{h} \neq \mathbf{0}$, we obtain $\mathbf{h} \cdot \mathbf{x}_n = \mathbf{h} \cdot (n\boldsymbol{\alpha}) = n(\mathbf{h} \cdot \boldsymbol{\alpha})$ for all $n \geq 1$. By the given linear independence property, $\mathbf{h} \cdot \boldsymbol{\alpha}$ is an irrational number, and so the sequence $(\mathbf{h} \cdot \mathbf{x}_n)_{n=1}^{\infty}$ is uniformly distributed modulo 1 by Theorem 4.1.10. It follows therefore from Corollary 4.1.31 that the sequence $(\mathbf{x}_n)_{n=1}^{\infty}$ is uniformly distributed modulo 1 in $\mathbb{R}^s$.

Now suppose that $1, \alpha_1, \ldots, \alpha_s$ are linearly dependent over $\mathbb{Q}$, say

$$r_1\alpha_1 + \cdots + r_s\alpha_s = r$$

with $r_1, \ldots, r_s, r \in \mathbb{Q}$ and at least one r_i, $1 \leq i \leq s$, different from 0. By clearing the denominators of $r_1, \ldots, r_s$, we deduce that $\mathbf{h}_0 \cdot \boldsymbol{\alpha} \in \mathbb{Q}$ for some $\mathbf{h}_0 \in \mathbb{Z}^s$ with $\mathbf{h}_0 \neq \mathbf{0}$. From $\mathbf{h}_0 \cdot \mathbf{x}_n = \mathbf{h}_0 \cdot (n\boldsymbol{\alpha}) = n(\mathbf{h}_0 \cdot \boldsymbol{\alpha})$ for all $n \geq 1$ and Theorem 4.1.10 we infer that the sequence $(\mathbf{h}_0 \cdot \mathbf{x}_n)_{n=1}^{\infty}$ is not uniformly distributed modulo 1, and so Corollary 4.1.31 shows that the sequence $(\mathbf{x}_n)_{n=1}^{\infty}$ is not uniformly distributed modulo 1 in $\mathbb{R}^s$. □

For $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_s) \in \mathbb{R}^s$ with $1, \alpha_1, \ldots, \alpha_s$ linearly independent over $\mathbb{Q}$, the sequence $(\{n\boldsymbol{\alpha}\})_{n=1}^{\infty}$ of fractional parts is called an s-dimensional *Kronecker sequence.*

Example 4.1.33 For a given dimension $s \geq 1$, let g be an irreducible polynomial over $\mathbb{Q}$ of degree $s + 1$ which has a real root α. Then for $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_s) \in \mathbb{R}^s$ with $\alpha_i = \alpha^i$ for $1 \leq i \leq s$, the irreducibility of g over $\mathbb{Q}$ implies immediately that $1, \alpha_1, \ldots, \alpha_s$ are linearly independent over $\mathbb{Q}$. Therefore the sequence $(n\boldsymbol{\alpha})_{n=1}^{\infty}$ is uniformly distributed modulo 1 in $\mathbb{R}^s$ by Theorem 4.1.32 and the sequence $(\{n\boldsymbol{\alpha}\})_{n=1}^{\infty}$ of fractional parts is an s-dimensional Kronecker sequence.

Example 4.1.34 Recall that an integer $m \geq 2$ is called squarefree if it is a product of distinct prime numbers. Now, for a given integer $s \geq 2$, we choose squarefree integers $m_1, \ldots, m_s$ that are pairwise coprime. We claim that $1, \sqrt{m_1}, \ldots, \sqrt{m_s}$ are linearly independent over $\mathbb{Q}$. We show even more, namely that

$$\sqrt{m_i} \notin \mathbb{Q}(\sqrt{m_1}, \ldots, \sqrt{m_{i-1}}) \qquad \text{for } 1 \leq i \leq s, \tag{4.17}$$

where $\mathbb{Q}(\sqrt{m_1}, \ldots, \sqrt{m_{i-1}})$ is the smallest subfield of $\mathbb{R}$ containing $\mathbb{Q}, \sqrt{m_1}, \ldots, \sqrt{m_{i-1}}$ (compare with Subsection 1.4.3). We proceed by induction on i. For $i = 1$ we have to verify that $\sqrt{m_i} = \sqrt{m_1} \notin \mathbb{Q}(\sqrt{m_1}, \ldots, \sqrt{m_{i-1}}) = \mathbb{Q}$, but this is a trivial fact. Suppose that we have proved (4.17) for some i with $1 \le i \le s-1$ and any pairwise coprime squarefree integers $m_1, \ldots, m_i$. Now we consider $i+1$ and we assume, on the contrary, that $\sqrt{m_{i+1}} \in \mathbb{Q}(\sqrt{m_1}, \ldots, \sqrt{m_i}) = F(\sqrt{m_i})$ with F being the field $\mathbb{Q}(\sqrt{m_1}, \ldots, \sqrt{m_{i-1}})$. Then we can write $\sqrt{m_{i+1}} = \theta_1 + \theta_2\sqrt{m_i}$ with $\theta_1, \theta_2 \in F$. If we had $\theta_1 = 0$, then $\sqrt{m_{i+1}m_i} = \theta_2 m_i \in F$, a contradiction to the induction hypothesis (4.17) applied with the squarefree integer $m_{i+1}m_i$ instead of m_i. If we had $\theta_2 = 0$, then $\sqrt{m_{i+1}} = \theta_1 \in F$, again a contradiction to (4.17). Thus $\theta_1\theta_2 \neq 0$, and so by squaring the identity $\sqrt{m_{i+1}} = \theta_1 + \theta_2\sqrt{m_i}$ we obtain $\sqrt{m_i} = (2\theta_1\theta_2)^{-1}(m_{i+1} - \theta_1^2 - \theta_2^2 m_i) \in F$, another contradiction to (4.17). The proof of (4.17) by induction is now complete. As we have already observed, this implies that $1, \sqrt{m_1}, \ldots, \sqrt{m_s}$ are linearly independent over $\mathbb{Q}$. With $\boldsymbol{\alpha} = (\sqrt{m_1}, \ldots, \sqrt{m_s}) \in \mathbb{R}^s$, the sequence $(n\boldsymbol{\alpha})_{n=1}^{\infty}$ is uniformly distributed modulo 1 in $\mathbb{R}^s$ by Theorem 4.1.32 and the sequence $(\{n\boldsymbol{\alpha}\})_{n=1}^{\infty}$ of fractional parts is an s-dimensional Kronecker sequence. The special case where $m_1, \ldots, m_s$ are distinct prime numbers is often considered in practice.

We have seen in Subsection 4.1.1 that the (star) discrepancy plays a crucial role in error bounds for one-dimensional quasi-Monte Carlo integration. The same holds true in the multidimensional case. For a point set $\mathcal{P}$ consisting of N points $\mathbf{x}_1, \ldots, \mathbf{x}_N$ in $[0,1)^s$ and every subinterval J of $[0,1]^s$, we write

$$A(J; \mathcal{P}) = \sum_{n=1}^{N} c_J(\mathbf{x}_n),$$

that is, $A(J;\mathcal{P})$ is the number of integers n with $1 \le n \le N$ for which $\mathbf{x}_n \in J$.

Definition 4.1.35 For a point set $\mathcal{P}$ consisting of N points in $[0,1)^s$, the (*extreme*) *discrepancy* $D_N(\mathcal{P})$ of $\mathcal{P}$ is defined by

$$D_N(\mathcal{P}) = \sup_J \left| \frac{A(J;\mathcal{P})}{N} - \lambda_s(J) \right|, \tag{4.18}$$

where the supremum is extended over all intervals $J = \prod_{i=1}^{s}[y_i, z_i)$ with $0 \le y_i < z_i \le 1$ for $1 \le i \le s$. If the supremum is extended only over the intervals J with $y_i = 0$ for $1 \le i \le s$, then we obtain the *star discrepancy* $D_N^*(\mathcal{P})$ of $\mathcal{P}$. For an infinite sequence $\mathcal{S}$ of points in $[0,1)^s$, we write $D_N(\mathcal{S})$, respectively $D_N^*(\mathcal{S})$, for the discrepancy, respectively star discrepancy, of the point set consisting of the first N terms of $\mathcal{S}$.

The s-dimensional generalization of Proposition 4.1.12 says that

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

for every point set $\mathcal{P}$ consisting of N points in $[0,1)^s$. In analogy with Theorem 4.1.14, the following criterion for uniform distribution of sequences holds in dimension s (see [41, Theorem 1.6 and Lemma 1.7]).

Theorem 4.1.36 *The following properties of a sequence $\mathcal{S}$ of points in $[0,1)^s$ are equivalent:*
(i) *$\mathcal{S}$ is uniformly distributed in $[0,1)^s$;*
(ii) *$\lim_{N\to\infty} D_N^*(\mathcal{S}) = 0$;*
(iii) *$\lim_{N\to\infty} D_N(\mathcal{S}) = 0$.*

Remark 4.1.37 There is a simple projection principle for (star) discrepancies. For dimensions r and s with $1 \leq r < s$, consider a projection $\pi_{s,r} : [0,1)^s \to [0,1)^r$ onto a subset of r coordinates. By using a suitable permutation of coordinates, we can assume without loss of generality that the projection is onto the first r coordinates. Thus, for $\mathbf{y} = (y_1, \ldots, y_s) \in [0,1)^s$ we define

$$\pi_{s,r}(\mathbf{y}) = (y_1, \ldots, y_r) \in [0,1)^r.$$

Let $\mathcal{P}$ be the point set comprising the points $\mathbf{x}_1, \ldots, \mathbf{x}_N \in [0,1)^s$ and let $\mathcal{P}^{(r)}$ be the point set consisting of the projected points $\pi_{s,r}(\mathbf{x}_1), \ldots, \pi_{s,r}(\mathbf{x}_N) \in [0,1)^r$. Let $J^{(r)} \subseteq [0,1)^r$ be an interval occurring in the definition of $D_N(\mathcal{P}^{(r)})$ according to (4.18) and put $J = J^{(r)} \times [0,1)^{s-r} \subseteq [0,1)^s$. Then for every $\mathbf{y} \in [0,1)^s$, it is clear that $\pi_{s,r}(\mathbf{y}) \in J^{(r)}$ if and only if $\mathbf{y} \in J$, and so $A(J^{(r)}; \mathcal{P}^{(r)}) = A(J; \mathcal{P})$. It follows that

$$\left|\frac{A(J^{(r)}; \mathcal{P}^{(r)})}{N} - \lambda_r(J^{(r)})\right| = \left|\frac{A(J; \mathcal{P})}{N} - \lambda_s(J)\right| \leq D_N(\mathcal{P}),$$

and by forming the supremum over all intervals $J^{(r)}$ on the left-hand side we obtain $D_N(\mathcal{P}^{(r)}) \leq D_N(\mathcal{P})$. In the same way it is shown that $D_N^*(\mathcal{P}^{(r)}) \leq D_N^*(\mathcal{P})$. In particular, every s-dimensional (star) discrepancy is bounded from below by a corresponding one-dimensional (star) discrepancy. It follows therefore from Proposition 4.1.19 that always $D_N(\mathcal{P}) \geq 1/N$ and from Corollary 4.1.17 that always $D_N^*(\mathcal{P}) \geq 1/(2N)$.

Better lower bounds on $D_N(\mathcal{P})$ and $D_N^*(\mathcal{P})$ can be established for dimensions $s \geq 2$ by using sophisticated methods. A classical lower bound is due to Roth [176] and it says that

$$D_N(\mathcal{P}) \geq D_N^*(\mathcal{P}) \geq c_s N^{-1} (\log N)^{(s-1)/2} \tag{4.19}$$

for every point set $\mathcal{P}$ of N points in $[0,1)^s$, where c_s is a positive constant depending only on the dimension s. In the case $s = 2$, the factor $(\log N)^{1/2}$ can be replaced by $\log N$ according to a result of Schmidt [177], that is,

$$D_N(\mathcal{P}) \geq D_N^*(\mathcal{P}) \geq c N^{-1} \log N \tag{4.20}$$

for every point set $\mathcal{P}$ of N points in $[0,1)^2$ with an absolute constant $c > 0$. Proofs of these bounds can be found in the book of Kuipers and Niederreiter [91, Section 2.2]. Minor improvements on the exponent $(s-1)/2$ of $\log N$ have been obtained recently for $s \geq 3$; we refer to the book of Dick and Pillichshammer [39, Section 3.2] for a survey of these improvements.

These results clearly imply lower bounds on the (star) discrepancy of infinite sequences. For instance, (4.19) shows that every sequence $\mathcal{S}$ of points in $[0,1)^s$ satisfies $D_N^*(\mathcal{S}) \geq c_s N^{-1}(\log N)^{(s-1)/2}$ for all $N \geq 1$. However, there is a simple trick based on the following lemma which allows us to establish a better lower bound for infinitely many N.

Lemma 4.1.38 *Let $s \geq 1$ and $N \geq 1$ be integers and let $\mathcal{S} = (\mathbf{x}_n)_{n=1}^{\infty}$ be a sequence of points in $[0,1)^s$. Let $\mathcal{P}$ be the point set consisting of the N points $((n-1)/N, \mathbf{x}_n) \in [0,1)^{s+1}$ for $n = 1, \ldots, N$. Then*

$$ND_N^*(\mathcal{P}) \leq \max_{1 \leq M \leq N} MD_M^*(\mathcal{S}) + 1.$$

Proof. For an arbitrary interval $J \subseteq [0,1)^{s+1}$ of the form $J = \prod_{i=1}^{s+1}[0, z_i)$, it is obvious that $((n-1)/N, \mathbf{x}_n) \in J$ if and only if $\mathbf{x}_n \in J' := \prod_{i=2}^{s+1}[0, z_i)$ and $n < Nz_1+1$. If M is the largest integer $< Nz_1 + 1$, then $A(J;\mathcal{P}) = A(J';\mathcal{P}_M)$, where $\mathcal{P}_M$ is the point set consisting of the first M terms of $\mathcal{S}$. Therefore

$$\begin{aligned}|A(J;\mathcal{P}) - N\lambda_{s+1}(J)| &\leq |A(J';\mathcal{P}_M) - M\lambda_s(J')| + |M\lambda_s(J') - N\lambda_{s+1}(J)| \\ &\leq MD_M^*(\mathcal{S}) + |M\lambda_s(J') - N\lambda_{s+1}(J)|.\end{aligned}$$

Now $Nz_1 \leq M < Nz_1 + 1$, hence

$$0 \leq M\lambda_s(J') - N\lambda_{s+1}(J) \leq (Nz_1 + 1)\prod_{i=2}^{s+1} z_i - N\prod_{i=1}^{s+1} z_i = \prod_{i=2}^{s+1} z_i \leq 1,$$

and the desired result follows. □

Theorem 4.1.39 *For every dimension $s \geq 1$ there exists a constant $c_s' > 0$, depending only on s, such that every sequence $\mathcal{S}$ of points in $[0,1)^s$ satisfies*

$$D_N(\mathcal{S}) \geq D_N^*(\mathcal{S}) \geq c_s' N^{-1}(\log N)^{s/2}$$

for infinitely many positive integers N.

Proof. By Lemma 4.1.38 and (4.19) (with s replaced by $s+1$), for every integer $N \geq 1$ there exists an integer M with $1 \leq M \leq N$ such that

$$MD_M^*(\mathcal{S}) \geq c_s(\log N)^{s/2} - 1.$$

Thus, for a suitable constant c_s' with $0 < c_s' < c_s$ and for sufficiently large N we get

$$MD_M^*(\mathcal{S}) \geq c_s'(\log N)^{s/2} \geq c_s'(\log M)^{s/2}.$$

It remains to prove that there are infinitely many values of M for which the last lower bound on $MD_M^*(\mathcal{S})$ holds. Suppose there were only finitely many possible choices for M and let M^* be the maximal choice. Then there exists a sufficiently large integer N with

$$c_s'(\log N)^{s/2} > \max_{1 \leq M \leq M^*} MD_M^*(\mathcal{S}). \tag{4.21}$$

For this N there exists again an integer M_1 with $1 \leq M_1 \leq N$ such that

$$M_1 D_{M_1}^*(\mathcal{S}) \geq c_s'(\log N)^{s/2} \geq c_s'(\log M_1)^{s/2}.$$

We must have $M_1 \leq M^*$ by the definition of M^*, and so (4.21) implies that $M_1 D_{M_1}^*(\mathcal{S}) < c_s'(\log N)^{s/2}$, which is a contradiction. □

Theorem 4.1.40 *There exists an absolute constant $c' > 0$ such that every sequence $\mathcal{S}$ of points in $[0,1)$ satisfies*

$$D_N(\mathcal{S}) \geq D_N^*(\mathcal{S}) \geq c' N^{-1} \log N$$

for infinitely many positive integers N.

Proof. This is shown in the same way as Theorem 4.1.39, but with the bound (4.20) instead of (4.19). □

To what extent are these lower bounds on the (star) discrepancy best possible? We will see in Subsection 4.2.2 that for every dimension $s \geq 1$ there is a construction of a sequence $\mathcal{S}$ of points in $[0,1)^s$ for which

$$D_N^*(\mathcal{S}) = O(N^{-1}(\log N)^s) \qquad \text{for all } N \geq 2, \tag{4.22}$$

where the implied constant is independent of N. An s-dimensional sequence $\mathcal{S}$ satisfying (4.22) is called a *low-discrepancy sequence*. We conclude that the lower bound in Theorem 4.1.40 for $s = 1$ is best possible. For $s \geq 2$ there is a gap in the exponent of $\log N$ when one compares the lower bound in Theorem 4.1.39 and the upper bound in (4.22). The determination of the best possible exponent of $\log N$ for $s \geq 2$ is *the* big open problem in discrepancy theory.

In the case $s = 1$ we have observed in Remark 4.1.18 that for every $N \geq 1$ we can easily construct a point set $\mathcal{P}$ of N points in $[0,1)$ that achieves the minimum value $D_N^*(\mathcal{P}) = 1/(2N)$ of the star discrepancy of any N points in $[0,1)$. For $s \geq 2$

it follows immediately from (4.22) and Lemma 4.1.38 that for every $N \geq 2$ we can construct a point set $\mathcal{P}$ of N points in $[0,1)^s$ for which

$$D_N^*(\mathcal{P}) = O(N^{-1}(\log N)^{s-1}) \tag{4.23}$$

with an implied constant independent of N (in fact, we like a constant depending only on s). For every $s \geq 1$, a point set $\mathcal{P}$ consisting of N points in $[0,1)^s$ and satisfying (4.23) is called a *low-discrepancy point set*. It is clear from what has been said above that the order of magnitude in (4.23) is best possible for $s = 1$ and $s = 2$. For $s \geq 3$ we run again into the big open problem of discrepancy theory concerning the best possible exponent of $\log N$, in this case the version for point sets.

We already mentioned one-dimensional quasi-Monte Carlo integration in Subsection 4.1.1. The full power of this approach is achieved in the multidimensional case where it can outperform Cartesian products of one-dimensional integration rules and even the Monte Carlo method. Numerical integration by means of multidimensional *quasi-Monte Carlo integration* uses the approximation

$$\int_{[0,1]^s} f(\mathbf{u})d\mathbf{u} \approx \frac{1}{N}\sum_{n=1}^{N} f(\mathbf{x}_n) \tag{4.24}$$

with nodes $\mathbf{x}_1, \ldots, \mathbf{x}_N \in [0,1)^s$. This looks formally like the Monte Carlo estimate (4.15), but the viewpoint is different: whereas (4.15) employs random samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$, the approximation (4.24) works with carefully chosen deterministic points $\mathbf{x}_1, \ldots, \mathbf{x}_N$. The underlying idea is that the Monte Carlo method captures the average performance of node sets, whereas in quasi-Monte Carlo integration we look for node sets that perform better than average. There is also a major difference in the analysis of these numerical integration techniques: the error bounds in the Monte Carlo method are probabilistic, whereas the error bounds for quasi-Monte Carlo integration are deterministic and typically involve some concept of discrepancy.

Quasi-Monte Carlo integration is an example of a *quasi-Monte Carlo method*, that is, a deterministic version of a Monte Carlo method. Quasi-Monte Carlo methods can be applied to other computational tasks, for instance, to optimization problems (see Section 4.5). Monte Carlo methods are an invention of the 1940s, as we already mentioned earlier, and multidimensional quasi-Monte Carlo integration was introduced shortly thereafter in the early 1950s. Indeed, the Los Alamos technical report of Richtmyer [168] from 1951 already coined the term "quasi-Monte Carlo method" and it proposed s-dimensional Kronecker sequences for quasi-Monte Carlo integration. In the one-dimensional case, the paper of Koksma [84] that established the Koksma inequality (see Theorem 4.1.21) can be considered a precursor of this work. Systematic research on quasi-Monte Carlo methods was begun in the Soviet Union in the late 1950s and the first book on quasi-Monte Carlo methods, namely that of Korobov [87], was published there in 1963. A comprehensive account of the work on quasi-Monte Carlo methods up to 1978 is presented in the survey article of Niederreiter [127]. More contemporary expository treatments of quasi-Monte Carlo methods can be found in the books of Dick and Pillichshammer [39], Leobacher and Pillichshammer [98], and Niederreiter [134]. It is remarkable that much of the basic research on quasi-Monte Carlo methods was carried out by number theorists.

Now we lead up to the standard error bound for multidimensional quasi-Monte Carlo integration which formally looks like Theorem 4.1.21, but where we have to be more careful about the definition of the variation $V(f)$. For a real-valued function f on $[0,1]^s$ and a subinterval J of $[0,1]^s$, let $\Delta(f;J)$ be an alternating sum of the values of f at the vertices of J (that is, function values at adjacent vertices have opposite signs). The *variation* of f on $[0,1]^s$ *in the sense of Vitali* is defined by

$$V^{(s)}(f) = \sup_{\mathcal{R}} \sum_{J \in \mathcal{R}} |\Delta(f;J)|,$$

where the supremum is extended over all partitions $\mathcal{R}$ of $[0,1]^s$ into subintervals. This is the straightforward generalization of the definition of the variation of a function on $[0,1]$. In the multidimensional case $s \geq 2$, we must take into account also the variation of projections of f since we encounter the phenomenon that $V^{(s)}(f) = 0$ if f depends on fewer than s variables. In detail, for integers $1 \leq k \leq s$ and $1 \leq i_1 < i_2 < \cdots < i_k \leq s$, let $V^{(k)}(f; i_1, \ldots, i_k)$ be the variation in the sense of Vitali of the restriction of f to the k-dimensional face

$$\{(u_1, \ldots, u_s) \in [0,1]^s : u_j = 1 \text{ for } j \neq i_1, \ldots, i_k\}$$

of the s-dimensional unit cube $[0,1]^s$. Then

$$V(f) := \sum_{k=1}^{s} \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq s} V^{(k)}(f; i_1, \ldots, i_k)$$

is called the *variation* of f on $[0,1]^s$ *in the sense of Hardy and Krause.* If $V(f)$ is finite, then we say that f has *bounded variation* on $[0,1]^s$ *in the sense of Hardy and Krause.* There is a useful sufficient condition for f to have bounded variation on $[0,1]^s$ in the sense of Hardy and Krause, namely that the partial derivative $\partial^s f / \partial u_1 \cdots \partial u_s$ exists and is continuous on $[0,1]^s$. For later use, we record the convenient formula for $V(f)$ in the two-dimensional case under this smoothness condition, namely

$$V(f) = \int_0^1 \int_0^1 \Big| \frac{\partial^2 f(u_1, u_2)}{\partial u_1 \partial u_2} \Big| du_1 du_2 + \int_0^1 \Big| \frac{df(u_1, 1)}{du_1} \Big| du_1 + \int_0^1 \Big| \frac{df(1, u_2)}{du_2} \Big| du_2. \quad (4.25)$$

With this notion of variation, the following inequality due to Hlawka [63] is valid in the multidimensional case (see also [91, Section 2.5] for a different proof).

Theorem 4.1.41 (Koksma-Hlawka Inequality) *If the real-valued function f has bounded variation $V(f)$ on $[0,1]^s$ in the sense of Hardy and Krause and $\mathbf{x}_1, \ldots, \mathbf{x}_N \in [0,1)^s$ are arbitrary, then*

$$\Big| \int_{[0,1]^s} f(\mathbf{u}) d\mathbf{u} - \frac{1}{N} \sum_{n=1}^{N} f(\mathbf{x}_n) \Big| \leq V(f) D_N^*(\mathcal{P}),$$

where $D_N^*(\mathcal{P})$ is the star discrepancy of the point set $\mathcal{P}$ consisting of $\mathbf{x}_1, \ldots, \mathbf{x}_N$.

In view of the Koksma-Hlawka inequality, the strategy in quasi-Monte Carlo integration is now evident: we have basically no control over the given integrand f, but we can choose the point set $\mathcal{P}$ so as to make the star discrepancy $D_N^*(\mathcal{P})$, and therefore the bound on the integration error, as small as possible. This suggests to choose $\mathcal{P}$ as a low-discrepancy point set in the sense of (4.23). Then in terms of the number $N \geq 2$ of nodes, the integration error is $O(N^{-1}(\log N)^{s-1})$ for integrands of bounded variation on $[0,1]^s$ in the sense of Hardy and Krause, which in the asymptotic regime is significantly smaller than the Monte Carlo error bound $O(N^{-1/2})$ in (4.16). Thus, we can expect that the quasi-Monte Carlo method outperforms the Monte Carlo method for many types of integrands, and this is borne out by numerical experiments and practical experience. For instance, the monetary values of various sophisticated financial instruments can be computed in real time by means of quasi-Monte Carlo integration, whereas the Monte Carlo method would take very much longer for this task (see Paskov and Traub [158] for a famous case study). Would number theorists of past generations have dreamed that number theory will one day become relevant on Wall Street?

The basic difference between the Monte Carlo method and the quasi-Monte Carlo method can be elucidated pictorially. We compare a plot of random points in the unit square $[0,1]^2$ with a plot of a low-discrepancy point set in $[0,1]^2$ (see Figure 4.4). The constellation of random points exhibits clusters (that is, points coming close together) and holes (that is, relatively large regions without points), and this is how it should be for truly random points. On the other hand, the points of a low-discrepancy point set avoid clusters and tend to fill the holes. Overall, random points show a somewhat chaotic behavior and low-discrepancy point sets display a pleasing regular pattern. Numerical integration of good quality seems to favor nodes with an equitable and nicely structured distribution on the integration domain, and so the quasi-Monte Carlo method is better geared to numerical integration than the Monte Carlo method.

In practice it can be convenient to have some flexibility in the choice of the number N of nodes. For instance, we may initially work with a moderately large value of N and decide later to increase N in order to achieve a higher accuracy in the computation of the given integral. From the viewpoint of efficiency, it is desirable to be able to reuse the previously computed function values in this scenario. This

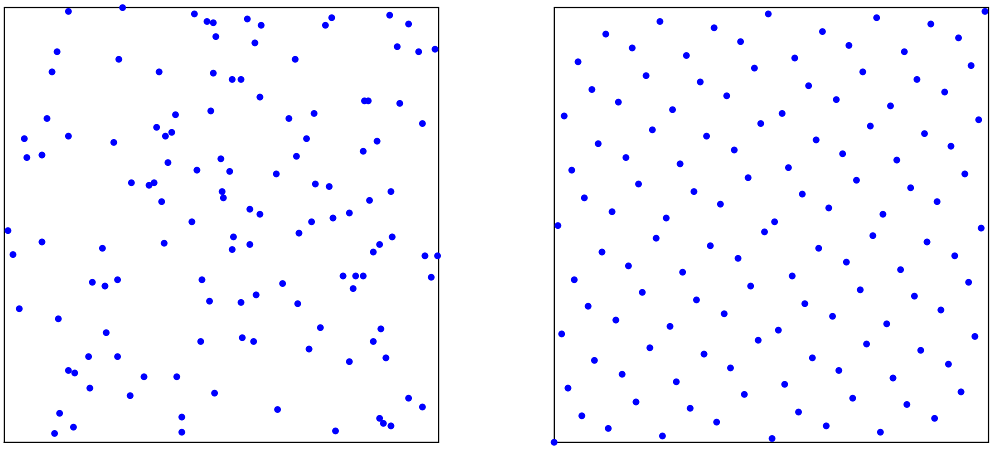

Figure 4.4: 128 random points (left) and the 128-element Hammersley point set in the base 2 (right)

suggests that we utilize a low-discrepancy sequence and take its first N terms as the integration nodes whenever a value of N has been selected. In this way, N can be increased while all data from an earlier computation with a smaller N can still be used. There is a relatively small price to pay for this convenience, namely the factor $\log N$ by which the discrepancy bounds (4.22) and (4.23) differ.

4.2 Classical low-discrepancy sequences

4.2.1 Kronecker sequences and continued fractions

When you read the previous section carefully, you realize that the remaining major issue in quasi-Monte Carlo integration is the construction of low-discrepancy point sets and sequences. In view of Lemma 4.1.38, we can focus on the construction of low-discrepancy sequences. In the one-dimensional case, Kronecker sequences $(\{n\alpha\})_{n=1}^{\infty}$ are low-discrepancy sequences for certain irrational numbers α. Suitable α can be determined by means of continued fractions, as we shall see below.

Continued fractions are a standard tool in number theory, and for the sake of convenience we review the basic facts about the continued fraction algorithm for irrational numbers. We start from an irrational number $\alpha = \alpha_0$, and further numbers $\alpha_1, \alpha_2, \ldots$ are obtained by the recursion

$$\alpha_{k+1} = \frac{1}{\{\alpha_k\}} \qquad \text{for } k = 0, 1, \ldots .$$

Note that all α_k are irrational, hence the fractional part $\{\alpha_k\}$ satisfies $\{\alpha_k\} \neq 0$, and so the definition of α_{k+1} makes sense. The *partial quotients* of α are defined by

$$a_k = \lfloor \alpha_k \rfloor \qquad \text{for } k = 0, 1, \ldots .$$

Then $a_0 = \lfloor \alpha \rfloor \in \mathbb{Z}$ and $a_k \in \mathbb{N}$ for $k \geq 1$. We can write

$$\alpha_k = a_k + \{\alpha_k\} = a_k + \frac{1}{\alpha_{k+1}} \qquad \text{for } k = 0, 1, \ldots . \tag{4.26}$$

By iterating the formula (4.26), we obtain

$$\alpha = \alpha_0 = a_0 + \frac{1}{\alpha_1} = a_0 + \frac{1}{a_1 + \frac{1}{\alpha_2}} = \cdots ,$$

and this leads to the infinite expansion

$$\alpha = a_0 + 1/(a_1 + 1/(a_2 + \cdots)) =: [a_0; a_1, a_2, \ldots] \tag{4.27}$$

called the *continued fraction expansion* of α. If for some $k \geq 0$ we terminate this expansion after the partial quotient a_k, then we get the *kth convergent* $r_k \in \mathbb{Q}$ to α. We write $r_k = p_k/q_k$ with $p_k, q_k \in \mathbb{Z}$ and $q_k \geq 1$. The numerators p_k and the denominators q_k can be computed by the recursions

$$\begin{aligned} p_{-2} &= 0,\ p_{-1} = 1,\ p_k = a_k p_{k-1} + p_{k-2} \ \text{ for } k \geq 0, \\ q_{-2} &= 1,\ q_{-1} = 0,\ q_k = a_k q_{k-1} + q_{k-2} \ \text{ for } k \geq 0. \end{aligned}$$

Note that $1 = q_0 \leq a_1 = q_1 < q_2 < \cdots$.

Lemma 4.2.1 *For all integers $k \geq -1$, the numerators and denominators of the convergents to α satisfy*

$$p_{k-1}q_k - p_k q_{k-1} = (-1)^k.$$

Proof. This is trivial for $k = -1$. Suppose that for some $k \geq 0$ the identity is shown for $k - 1$. Then

$$\begin{aligned} p_{k-1}q_k - p_k q_{k-1} &= p_{k-1}(a_k q_{k-1} + q_{k-2}) - (a_k p_{k-1} + p_{k-2})q_{k-1} \\ &= -(p_{k-2}q_{k-1} - p_{k-1}q_{k-2}) = -(-1)^{k-1} = (-1)^k, \end{aligned}$$

and the induction is complete. □

Lemma 4.2.1 implies that $\gcd(p_k, q_k) = 1$ for $k \geq 0$, and so p_k/q_k yields the rational number r_k in reduced form. We need two more facts about continued fractions.

Lemma 4.2.2 *The identity*

$$\alpha = \frac{p_k \alpha_{k+1} + p_{k-1}}{q_k \alpha_{k+1} + q_{k-1}}$$

holds for all integers $k \geq -1$.

Proof. We proceed again by induction on k. The formula is trivial for $k = -1$. Suppose that it is shown for some $k \geq -1$. Then by (4.26),

$$\begin{aligned} \alpha &= \frac{p_k \alpha_{k+1} + p_{k-1}}{q_k \alpha_{k+1} + q_{k-1}} = \frac{p_k\big(a_{k+1} + \frac{1}{\alpha_{k+2}}\big) + p_{k-1}}{q_k\big(a_{k+1} + \frac{1}{\alpha_{k+2}}\big) + q_{k-1}} \\ &= \frac{p_{k+1} + \frac{p_k}{\alpha_{k+2}}}{q_{k+1} + \frac{q_k}{\alpha_{k+2}}} = \frac{p_{k+1}\alpha_{k+2} + p_k}{q_{k+1}\alpha_{k+2} + q_k}, \end{aligned}$$

and so the formula holds for $k + 1$. □

Lemma 4.2.3 *The inequality*

$$|\alpha - r_k| < (q_k q_{k+1})^{-1}$$

holds for all integers $k \geq 0$.

Proof. By first applying Lemma 4.2.2 and then Lemma 4.2.1, we get

$$\begin{aligned} \alpha - r_k &= \frac{p_k \alpha_{k+1} + p_{k-1}}{q_k \alpha_{k+1} + q_{k-1}} - \frac{p_k}{q_k} \\ &= \frac{p_{k-1} q_k - p_k q_{k-1}}{q_k(q_k \alpha_{k+1} + q_{k-1})} = \frac{(-1)^k}{q_k(q_k \alpha_{k+1} + q_{k-1})}. \end{aligned}$$

Using $\alpha_{k+1} > \lfloor \alpha_{k+1} \rfloor = a_{k+1}$, we immediately obtain the desired inequality. □

Since $\lim_{k\to\infty} q_k = \infty$, Lemma 4.2.3 implies that $\lim_{k\to\infty} r_k = \alpha$. This justifies *a posteriori* the first identity in (4.27) and the terminology "kth convergent" for r_k.

We note the following simple principle pertaining to a superposition $\mathcal{P}$ of point sets $\mathcal{P}_1, \ldots, \mathcal{P}_m$, that is, a point set $\mathcal{P}$ obtained by listing in some order the points of $\mathcal{P}_1, \ldots, \mathcal{P}_m$ with the correct multiplicities. In the present subsection we need this principle only in the one-dimensional case, but it holds for any dimension.

Lemma 4.2.4 *Let $m \geq 1$ and $s \geq 1$ be integers. For $j = 1, \ldots, m$, let $\mathcal{P}_j$ be a point set of N_j points in $[0,1)^s$. Let $\mathcal{P}$ be the superposition of $\mathcal{P}_1, \ldots, \mathcal{P}_m$ which contains $N = \sum_{j=1}^m N_j$ points. Then*

$$N D_N(\mathcal{P}) \leq \sum_{j=1}^m N_j D_{N_j}(\mathcal{P}_j)$$

and also

$$N D_N^*(\mathcal{P}) \leq \sum_{j=1}^m N_j D_{N_j}^*(\mathcal{P}_j).$$

Proof. Let $J \subseteq [0,1)^s$ be an interval appearing in the supremum in (4.18). Then $A(J;\mathcal{P}) = \sum_{j=1}^{m} A(J;\mathcal{P}_j)$ by the definition of $\mathcal{P}$, and so

$$\begin{aligned}|A(J;\mathcal{P}) - N\lambda_s(J)| &= \Big|\sum_{j=1}^{m}(A(J;\mathcal{P}_j) - N_j\lambda_s(J))\Big| \\ &\leq \sum_{j=1}^{m}|A(J;\mathcal{P}_j) - N_j\lambda_s(J)| \leq \sum_{j=1}^{m} N_j D_{N_j}(\mathcal{P}_j).\end{aligned}$$

Taking the supremum on the left-hand side completes the proof of the first inequality. The inequality for the star discrepancy is shown similarly. □

We are now ready to establish a discrepancy bound for one-dimensional Kronecker sequences. We use the notation for continued fractions introduced above. In particular, the positive integers a_k, $k = 1, 2, \ldots$, denote partial quotients in the continued fraction expansion (4.27) of an irrational number α.

Theorem 4.2.5 *Let α be an irrational number and let $\mathcal{S} = (\{n\alpha\})_{n=1}^{\infty}$ be the corresponding Kronecker sequence. Every integer $N \geq 1$ can be represented in the form $N = \sum_{k=0}^{l(N)} c_k q_k$, where $l(N)$ is the unique nonnegative integer with $q_{l(N)} \leq N < q_{l(N)+1}$ and where the c_k are integers with $0 \leq c_k \leq a_{k+1}$ for $0 \leq k \leq l(N)$. Then*

$$ND_N(\mathcal{S}) < \sum_{\substack{k=0 \\ c_k \geq 1}}^{l(N)} (c_k + 1) \leq \sum_{k=1}^{l(N)+1} a_k.$$

Proof. Since $1 = q_0 \leq q_1 < q_2 < \cdots$, the existence and uniqueness of $l(N)$ is guaranteed. We can write $N = c_{l(N)} q_{l(N)} + d$ with integers $c_{l(N)} \geq 1$ and $0 \leq d < q_{l(N)}$. If we had $c_{l(N)} > a_{l(N)+1}$, then

$$N \geq c_{l(N)} q_{l(N)} \geq a_{l(N)+1} q_{l(N)} + q_{l(N)} \geq q_{l(N)+1},$$

a contradiction. Therefore $c_{l(N)} \leq a_{l(N)+1}$. If $d \geq 1$, then we apply this procedure to d instead of N and, continuing in this manner, we arrive at the desired representation for N.

Given this representation for N, we decompose the point set $\mathcal{P}$ consisting of the first N terms of $\mathcal{S}$ into blocks of consecutive terms, namely c_k blocks of length q_k for $0 \leq k \leq l(N)$. Of course, we need to consider only those k with $c_k \geq 1$. Take such a block of length q_k for a fixed k with $c_k \geq 1$; it is a point set $\mathcal{P}_k$ consisting of the fractional parts $\{n\alpha\}$, $n = n_k, n_k + 1, \ldots, n_k + q_k - 1$, for some integer $n_k \geq 1$. Let p_k/q_k be the kth convergent to α. Then on account of Lemma 4.2.3 we can write

$$\alpha = \frac{p_k}{q_k} + \frac{\delta_k}{q_k q_{k+1}} \qquad \text{with } |\delta_k| < 1.$$

Thus, if $n = n_k + j$, $j = 0, 1, \ldots, q_k - 1$, as above, then

$$\{n\alpha\} = \Big\{\frac{jp_k}{q_k} + n_k\alpha + \frac{j\delta_k}{q_k q_{k+1}}\Big\}.$$

Since $\gcd(p_k, q_k) = 1$, the fractional parts $\{jp_k/q_k + n_k\alpha\}$, $j = 0, 1, \ldots, q_k - 1$, form a point set $\mathcal{Q}_k$ of q_k equidistant points in $[0, 1)$ with distance $1/q_k$, and so $D_{q_k}(\mathcal{Q}_k) = 1/q_k$. Because

$$\Big|\frac{j\delta_k}{q_k q_{k+1}}\Big| < \frac{1}{q_{k+1}} \qquad \text{for } j = 0, 1, \ldots, q_k - 1,$$

the point set $\mathcal{P}_k$ is obtained by displacing modulo 1 the points of $\mathcal{Q}_k$ in one direction (which depends on the sign of δ_k) by distances $< 1/q_{k+1}$. Therefore

$$D_{q_k}(\mathcal{P}_k) < \frac{1}{q_k} + \frac{1}{q_{k+1}}.$$

From Lemma 4.2.4 and the way in which we decomposed $\mathcal{P}$, we obtain

$$ND_N(\mathcal{S}) = ND_N(\mathcal{P}) < \sum_{\substack{k=0\\ c_k \geq 1}}^{l(N)} c_k\Big(1 + \frac{q_k}{q_{k+1}}\Big) \leq \sum_{\substack{k=0\\ c_k \geq 1}}^{l(N)} \Big(c_k + \frac{a_{k+1}q_k}{q_{k+1}}\Big) \leq \sum_{\substack{k=0\\ c_k \geq 1}}^{l(N)} (c_k + 1),$$

which is the first bound for $ND_N(\mathcal{S})$ in the theorem.

If $c_0 \geq 1$, then in the last step of the algorithm at the beginning of the proof we have $d = c_1 q_1 + c_0$ with $1 \leq c_0 < q_1$, and so $c_0 + 1 \leq q_1 = a_1$. If $c_k = a_{k+1}$ for some $k \geq 1$, then we claim that $c_{k-1} = 0$. Indeed, if $q_k \leq d < q_{k+1}$ and $d = c_k q_k + d_1$ with $c_k = a_{k+1}$, then

$$d_1 = d - c_k q_k = d - a_{k+1}q_k < q_{k+1} - a_{k+1}q_k = q_{k-1},$$

and so $c_{k-1} = 0$. Using these properties of the c_k, we deduce the second bound for $ND_N(\mathcal{S})$ in the theorem from the first bound. □

The order of magnitude of the discrepancy bound in Theorem 4.2.5 depends on the size of the partial quotients of α. A particularly attractive case occurs if α has *bounded partial quotients*, which means that the partial quotients of α are uniformly bounded.

Theorem 4.2.6 *Let α be an irrational number for which there exists a positive integer K such that the partial quotients a_k of α satisfy $a_k \leq K$ for all $k \geq 1$. Then the corresponding Kronecker sequence $\mathcal{S} = (\{n\alpha\})_{n=1}^{\infty}$ satisfies the discrepancy bound*

$$D_N(\mathcal{S}) < G(K)N^{-1}\log(N+1) \qquad \textit{for all } N \geq 1,$$

where $G(K) = (K+1)/\log(K+1)$ for $K \neq 2$ and $G(2) = 2/\log 2$.

Proof. In view of Theorem 4.2.5, it suffices to show that

$$s(N) := \sum_{\substack{k=0 \\ c_k \geq 1}}^{l(N)} (c_k + 1) \leq G(K) \log(N+1) \qquad \text{for all } N \geq 1. \tag{4.28}$$

Here $s(N)$ is well defined if we use the coefficients c_k produced by the algorithm at the beginning of the proof of Theorem 4.2.5. We formally put $s(0) = 0$.

We establish (4.28) by induction on the value of $l(N)$. If $q_0 < q_1$, then the least possible value of $l(N)$ is 0 and a corresponding N satisfies $1 \leq N < q_1 = a_1 \leq K$. If $q_0 = q_1 = 1$, then the least possible value of $l(N)$ is 1 and a corresponding N satisfies $1 \leq N \leq q_2 - 1 = a_2 \leq K$. Since $s(N) = N + 1$ for these N, it suffices to verify for the first step in the induction that

$$N + 1 \leq G(K) \log(N+1) \qquad \text{for } 1 \leq N \leq K. \tag{4.29}$$

But this follows from the fact that $G(K) = \max_{1 \leq u \leq K} (u+1)/\log(u+1)$.

Now we consider an arbitrary l with $q_l > 1$ and a corresponding N with $l(N) = l$, hence with $q_l \leq N < q_{l+1}$. We write $N = c_l q_l + d$ with $0 \leq d < q_l$. Then $s(N) = c_l + 1 + s(d)$, and the induction hypothesis yields

$$s(N) \leq c_l + 1 + G(K) \log(d+1),$$

which holds also for $d = 0$. Now $N + 1 = c_l q_l + d + 1 \geq (c_l + 1)(d + 1)$ and $1 \leq c_l \leq a_{l+1} \leq K$. Thus by (4.29),

$$s(N) \leq G(K) \log(c_l + 1) + G(K) \log(d+1) \leq G(K) \log(N+1)$$

and (4.28) is shown by induction. □

Example 4.2.7 A famous example of an irrational number with bounded partial quotients is $\alpha = (\sqrt{5} - 1)/2 = 0.618\ldots$, or one may also take the golden ratio $\alpha + 1 = (\sqrt{5} + 1)/2$. Note that α satisfies $\alpha^2 + \alpha = 1$. Here $a_0 = \lfloor \alpha \rfloor = 0$. Next we get $\alpha_1 = \alpha^{-1} = \alpha + 1$, and so $\{\alpha_1\} = \alpha$. Therefore in the next step $\alpha_2 = \alpha^{-1} = \alpha + 1$ and $\{\alpha_2\} = \alpha$. Hence it is clear that $\alpha_k = \alpha + 1$ and $a_k = \lfloor \alpha_k \rfloor = 1$ for all $k \geq 1$. Thus, α has the periodic continued fraction expansion

$$\alpha = [0; 1, 1, 1, \ldots].$$

The Kronecker sequence $\mathcal{S} = (\{n\alpha\})_{n=1}^{\infty} = (\{n(\sqrt{5} - 1)/2\})_{n=1}^{\infty}$ is a low-discrepancy sequence with

$$D_N(\mathcal{S}) < \frac{2}{\log 2} N^{-1} \log(N+1) \qquad \text{for all } N \geq 1$$

by Theorem 4.2.6. More generally, if β is any quadratic irrational, then β has a periodic continued fraction expansion according to a classical theorem of Lagrange (see [172, Section III.1] for two different proofs of this result), and so β has bounded partial quotients. Therefore $(\{n\beta\})_{n=1}^{\infty}$ is also a low-discrepancy sequence. If you want to learn more about the beautiful theory of continued fractions, we refer again to the book of Rockett and Szűsz [172].

In the multidimensional case, the theory of Kronecker sequences is much less satisfactory. There is of course the criterion in Theorem 4.1.32 for Kronecker sequences, but it is much harder to get strong discrepancy bounds for multidimensional Kronecker sequences, mainly because it is not known how to design a multidimensional continued fraction algorithm that is every bit as good as the one-dimensional continued fraction algorithm. A probabilistic result due to Beck [9] says the following: for a given dimension $s \geq 1$, pick a point $\boldsymbol{\alpha}$ at random from the probability space $[0,1]^s$ supplied with the s-dimensional Lebesgue measure; then with probability 1 the sequence $\mathcal{S}(\boldsymbol{\alpha}) := (\{n\boldsymbol{\alpha}\})_{n=1}^{\infty}$ is a Kronecker sequence and for every $\varepsilon > 0$ its discrepancy satisfies

$$D_N(\mathcal{S}(\boldsymbol{\alpha})) = O(N^{-1}(\log N)^s(\log\log N)^{1+\varepsilon}) \qquad \text{for all } N \geq 3,$$

where the implied constant depends only on ε and $\boldsymbol{\alpha}$. Thus with probability 1, Kronecker sequences are "almost" low-discrepancy sequences in the sense of (4.22). However, for $s \geq 2$ not a single *explicit* $\boldsymbol{\alpha}$ is known for which the above discrepancy bound for $\mathcal{S}(\boldsymbol{\alpha})$ holds. There is a weaker deterministic result for an interesting family of Kronecker sequences: if $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_s) \in \mathbb{R}^s$ with algebraic numbers $\alpha_1, \ldots, \alpha_s$ such that $1, \alpha_1, \ldots, \alpha_s$ are linearly independent over $\mathbb{Q}$, then for every $\varepsilon > 0$ the discrepancy bound $D_N(\mathcal{S}(\boldsymbol{\alpha})) = O(N^{-1+\varepsilon})$ holds for all $N \geq 1$, where the implied constant depends only on ε and $\boldsymbol{\alpha}$ (see [124]). This discrepancy bound applies, for instance, to the points $\boldsymbol{\alpha}$ constructed in Examples 4.1.33 and 4.1.34.

4.2.2 Halton sequences

Historically the first construction of low-discrepancy sequences for arbitrary dimensions was devised by Halton [58] in 1960 and it is based on elementary number theory. For an integer $b \geq 2$, we again write $Z_b = \{0, 1, \ldots, b-1\}$ for the least residue system modulo b. Every integer $n \geq 0$ has a unique digit expansion

$$n = \sum_{j=0}^{\infty} z_j(n) b^j \tag{4.30}$$

in base b, where $z_j(n) \in Z_b$ for all $j \geq 0$ and $z_j(n) = 0$ for all sufficiently large j. The *radical-inverse function* ϕ_b in base b is defined by

$$\phi_b(n) = \sum_{j=0}^{\infty} z_j(n) b^{-j-1} \in [0,1) \qquad \text{for } n = 0, 1, \ldots.$$

Since the set of nonnegative integers is the natural domain of radical-inverse functions, it is reasonable to commence the enumeration of the terms of sequences derived from radical-inverse functions with the index $n = 0$.

Definition 4.2.8 For a dimension $s \geq 1$ and integers $b_1, \ldots, b_s \geq 2$ that are pairwise coprime if $s \geq 2$, the *Halton sequence in the bases* $b_1, \ldots, b_s$ is the sequence $(\mathbf{x}_n)_{n=0}^{\infty}$ with

$$\mathbf{x}_n = (\phi_{b_1}(n), \ldots, \phi_{b_s}(n)) \in [0,1)^s \qquad \text{for } n = 0, 1, \ldots.$$

Remark 4.2.9 In the case $s = 1$ and with $b_1 = b$, the sequence $(\phi_b(n))_{n=0}^{\infty}$ is called the *van der Corput sequence in base* b. This one-dimensional low-discrepancy sequence was already introduced, at least for the base $b = 2$, by van der Corput [28] in 1935.

Example 4.2.10 We compute the first eight terms of the van der Corput sequence $(\phi_2(n))_{n=0}^{\infty}$ in base 2. In the table below, we first list n in its decimal form, then n in binary, then $\phi_2(n)$ in binary, and finally $\phi_2(n)$ as a rational number in reduced form.

n	0	1	2	3	4	5	6	7
binary n	000	001	010	011	100	101	110	111
binary $\phi_2(n)$	0.000	0.100	0.010	0.110	0.001	0.101	0.011	0.111
$\phi_2(n)$	0	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{1}{8}$	$\frac{5}{8}$	$\frac{3}{8}$	$\frac{7}{8}$

For the proof of the property that every Halton sequence is a low-discrepancy sequence (see Theorem 4.2.14 below), we need several auxiliary results.

Lemma 4.2.11 *Let $b \geq 2$ and $n \geq 0$ be integers and let v and f be positive integers with $v \leq b^f$. Then $\phi_b(n) \in [0, vb^{-f})$ if and only if $n \in \cup_{k=1}^{h} Q_k$, where $1 \leq h \leq bf$, each Q_k is a residue class in $\mathbb{Z}$ with modulus m_k, the residue classes $Q_1, \ldots, Q_h$ are disjoint and independent of n, and $\sum_{k=1}^{h} m_k^{-1} = vb^{-f}$.*

Proof. We write $(v-1)b^{-f} = \sum_{j=0}^{f-1} d_j b^{-j-1}$ with $d_j \in Z_b$ for $0 \leq j \leq f-1$. Then $\phi_b(n) \in [0, vb^{-f})$ if and only if

$$\sum_{j=0}^{f-1} z_j(n) b^{-j-1} \leq \sum_{j=0}^{f-1} d_j b^{-j-1},$$

with the notation in (4.30). This condition holds if and only if one of the following f mutually exclusive conditions is satisfied: (C_1) $z_0(n) \leq d_0 - 1$; (C_2) $z_0(n) = d_0$ and $z_1(n) \leq d_1 - 1$; (C_3) $z_0(n) = d_0$, $z_1(n) = d_1$, and $z_2(n) \leq d_2 - 1$;...; (C_f) $z_0(n) = d_0, \ldots, z_{f-2}(n) = d_{f-2}$, and $z_{f-1}(n) \leq d_{f-1}$. These conditions can be translated into the following congruence conditions on n: (C_1') $n \equiv g_0 \pmod{b}$ for some $0 \le g_0 \le d_0 - 1$; (C_2') $n \equiv d_0 + g_1 b \pmod{b^2}$ for some $0 \le g_1 \le d_1 - 1$; (C_3') $n \equiv d_0 + d_1 b + g_2 b^2 \pmod{b^3}$ for some $0 \le g_2 \le d_2 - 1$;...; (C_f') $n \equiv d_0 + d_1 b + \cdots + d_{f-2} b^{f-2} + g_{f-1} b^{f-1} \pmod{b^f}$ for some $0 \le g_{f-1} \le d_{f-1}$. This yields disjoint residue classes $Q_1, \dots, Q_h$ in which n must lie. The number h of residue classes satisfies

$$h = \sum_{j=0}^{f-2} d_j + d_{f-1} + 1 \le (b-1)f + 1 \le bf.$$

As to the moduli $m_1, \dots, m_h$ of $Q_1, \dots, Q_h$, respectively, we have d_0 moduli equal to b, d_1 moduli equal to b^2,..., d_{f-2} moduli equal to b^{f-1}, and $d_{f-1}+1$ moduli equal to b^f. Therefore

$$\sum_{k=1}^{h} m_k^{-1} = \sum_{j=0}^{f-2} d_j b^{-j-1} + (d_{f-1}+1) b^{-f} = (v-1)b^{-f} + b^{-f} = vb^{-f},$$

and all assertions are shown. □

Lemma 4.2.12 *For a dimension $s \ge 1$, let $b_1, \dots, b_s \ge 2$ be integers that are pairwise coprime if $s \ge 2$ and let $n \ge 0$ be an integer. Let $v_1, \dots, v_s$ and $f_1, \dots, f_s$ be positive integers with $v_i \le b_i^{f_i}$ for $1 \le i \le s$. Then*

$$\mathbf{x}_n = (\phi_{b_1}(n), \dots, \phi_{b_s}(n)) \in J := \prod_{i=1}^{s} [0, v_i b_i^{-f_i})$$

if and only if $n \in \cup_{k=1}^{H} R_k$, where $1 \le H \le b_1 \cdots b_s f_1 \cdots f_s$, each R_k is a residue class in $\mathbb{Z}$ with modulus m_k, the residue classes $R_1, \dots, R_H$ are disjoint and independent of n, and $\sum_{k=1}^{H} m_k^{-1} = \lambda_s(J)$.

Proof. The case $s = 1$ was proved in Lemma 4.2.11, and so we can assume that $s \ge 2$. Note that $\mathbf{x}_n \in J$ if and only if $\phi_{b_i}(n) \in [0, v_i b_i^{-f_i})$ for $1 \le i \le s$. For each fixed $i = 1, \dots, s$, we apply Lemma 4.2.11 and this yields the condition $n \in \cup_{k=1}^{h_i} Q_k^{(i)}$ for disjoint residue classes $Q_1^{(i)}, \dots, Q_{h_i}^{(i)}$ with respective moduli $m_1^{(i)}, \dots, m_{h_i}^{(i)}$. Furthermore, $h_i \le b_i f_i$ and $\sum_{k=1}^{h_i} (m_k^{(i)})^{-1} = v_i b_i^{-f_i}$ for $1 \le i \le s$. Since $b_1, \dots, b_s$ are pairwise coprime, we can combine these conditions for $i = 1, \dots, s$ by the Chinese remainder theorem to arrive at the condition $n \in \cup_{k=1}^{H} R_k$ in the lemma, where $1 \le H = h_1 \cdots h_s \le b_1 \cdots b_s f_1 \cdots f_s$. Furthermore, the new moduli $m_1, \dots, m_H$ are exactly all products $m_{k_1}^{(1)} \cdots m_{k_s}^{(s)}$ with $1 \le k_i \le h_i$ for $1 \le i \le s$. Therefore

$$\sum_{k=1}^{H} m_k^{-1} = \prod_{i=1}^{s} \Big(\sum_{k=1}^{h_i} (m_k^{(i)})^{-1} \Big) = \prod_{i=1}^{s} v_i b_i^{-f_i} = \lambda_s(J),$$

as claimed. □

Lemma 4.2.13 *If $u_i, w_i \in [0,1]$ for $1 \le i \le s$, then*

$$\Big|\prod_{i=1}^{s} u_i - \prod_{i=1}^{s} w_i\Big| \le \sum_{i=1}^{s} |u_i - w_i|.$$

Proof. We proceed by induction on s. The case $s = 1$ is trivial. If the inequality is shown for some $s \ge 1$, then

$$\begin{aligned}\Big|\prod_{i=1}^{s+1} u_i - \prod_{i=1}^{s+1} w_i\Big| &= \Big|(u_{s+1} - w_{s+1})\prod_{i=1}^{s} u_i + w_{s+1}\Big(\prod_{i=1}^{s} u_i - \prod_{i=1}^{s} w_i\Big)\Big| \\ &\le |u_{s+1} - w_{s+1}| + w_{s+1}\sum_{i=1}^{s} |u_i - w_i| \le \sum_{i=1}^{s+1} |u_i - w_i|,\end{aligned}$$

and the induction is complete. □

Theorem 4.2.14 *Let $s \ge 1$ be a given dimension and let $b_1, \dots, b_s \ge 2$ be integers that are pairwise coprime if $s \ge 2$. Then the star discrepancy of the Halton sequence $\mathcal{S}$ in the bases $b_1, \dots, b_s$ satisfies*

$$D_N^*(\mathcal{S}) \le C(b_1, \dots, b_s) N^{-1} (\log N)^s \qquad \textit{for all } N \ge 2$$

with a constant $C(b_1, \dots, b_s) > 0$ depending only on $b_1, \dots, b_s$.

Proof. We fix $N \ge 2$ and let $\mathcal{P}_N$ be the point set consisting of the first N terms $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{N-1}$ of $\mathcal{S}$. We introduce the positive integers

$$f_i = \left\lceil \frac{\log N}{\log b_i} \right\rceil \qquad \text{for } 1 \le i \le s. \tag{4.31}$$

We first consider an interval $J \subseteq [0,1)^s$ of the form

$$J = \prod_{i=1}^{s} [0, v_i b_i^{-f_i})$$

with integers $v_1, \dots, v_s$ satisfying $1 \le v_i \le b_i^{f_i}$ for $1 \le i \le s$. Then applying Lemma 4.2.12 and its notation, we obtain

$$A(J; \mathcal{P}_N) = \sum_{k=1}^{H} B_N(R_k),$$

where $B_N(R_k)$ is the number of integers n with $0 \le n \le N-1$ lying in the residue class R_k with modulus m_k. Since any m_k consecutive integers contain exactly one element of R_k, we can write $B_N(R_k) = \lfloor N/m_k \rfloor + \theta_N(k)$ with $\theta_N(k)$ being either 0 or 1, and so $B_N(R_k) = N/m_k + \beta_N(k)$ with $|\beta_N(k)| \le 1$. It follows that

$$\begin{aligned} |A(J;\mathcal{P}_N) - N\lambda_s(J)| &= \Big| \sum_{k=1}^{H} \Big(\frac{N}{m_k} + \beta_N(k) \Big) - N \sum_{k=1}^{H} \frac{1}{m_k} \Big| \\ &= \Big| \sum_{k=1}^{H} \beta_N(k) \Big| \le H \le b_1 \cdots b_s f_1 \cdots f_s \end{aligned}$$

by Lemma 4.2.12. This bound holds trivially if some v_i are 0, that is, if J is empty.

Now we consider an arbitrary interval $J = \prod_{i=1}^{s} [0, w_i) \subseteq [0,1)^s$ appearing in the definition of the star discrepancy. We choose integers $v_1, \ldots, v_s$ such that $(v_i - 1)b_i^{-f_i} \le w_i \le v_i b_i^{-f_i}$ and $1 \le v_i \le b_i^{f_i}$ for $1 \le i \le s$. We introduce the intervals

$$J_1 = \prod_{i=1}^{s} [0, (v_i - 1)b_i^{-f_i}), \quad J_2 = \prod_{i=1}^{s} [0, v_i b_i^{-f_i}).$$

Then $J_1 \subseteq J \subseteq J_2$, and so

$$\begin{aligned} &A(J_1;\mathcal{P}_N) - N\lambda_s(J_1) + N(\lambda_s(J_1) - \lambda_s(J_2)) \\ &\le A(J;\mathcal{P}_N) - N\lambda_s(J) \\ &\le A(J_2;\mathcal{P}_N) - N\lambda_s(J_2) + N(\lambda_s(J_2) - \lambda_s(J_1)). \end{aligned}$$

By what we have already shown for the intervals J_1 and J_2, we get

$$|A(J;\mathcal{P}_N) - N\lambda_s(J)| \le b_1 \cdots b_s f_1 \cdots f_s + N(\lambda_s(J_2) - \lambda_s(J_1)),$$

and an application of Lemma 4.2.13 yields

$$ND_N^*(\mathcal{S}) = ND_N^*(\mathcal{P}_N) \le b_1 \cdots b_s f_1 \cdots f_s + N \sum_{i=1}^{s} b_i^{-f_i}.$$

By using the definition of $f_1, \ldots, f_s$ in (4.31), we arrive at the final result. □

An explicit form of the discrepancy bound in Theorem 4.2.14 can be found in [39, Theorem 3.36]. Typically, the constant $C(b_1, \ldots, b_s)$ in Theorem 4.2.14 becomes smaller for smaller values of the bases $b_1, \ldots, b_s$. Thus, a popular choice is to take $b_1, \ldots, b_s$ as the first s prime numbers, that is, $b_1 = 2$, $b_2 = 3$, $b_3 = 5$, and so on. We observe that every s-dimensional Halton sequence is uniformly distributed in $[0,1)^s$ by Theorems 4.1.36 and 4.2.14.

For dimensions $s \ge 2$, we can construct low-discrepancy point sets in the sense of (4.23) by using Halton sequences and the idea in Lemma 4.1.38. Let $b_1, \ldots, b_{s-1} \ge 2$ be integers that are pairwise coprime if $s \ge 3$. For an integer $N \ge 2$, let $\mathcal{P}$ be the point set consisting of the points

$$\mathbf{y}_n = \Big(\frac{n}{N}, \phi_{b_1}(n), \ldots, \phi_{b_{s-1}}(n) \Big) \in [0,1)^s \qquad \text{for } n = 0, 1, \ldots, N-1. \tag{4.32}$$

Such a point set $\mathcal{P}$ is called a *Hammersley point set*, after the work of Hammersley [59]. If we want to stress the role of the integers $b_1, \ldots, b_{s-1}$ in this construction, then we speak of a Hammersley point set in the bases $b_1, \ldots, b_{s-1}$.

Theorem 4.2.15 *The star discrepancy of the Hammersley point set $\mathcal{P}$ in* (4.32) *satisfies*

$$D_N^*(\mathcal{P}) \leq C_1(b_1, \ldots, b_{s-1}) N^{-1} (\log N)^{s-1}$$

with a constant $C_1(b_1, \ldots, b_{s-1}) > 0$ depending only on $b_1, \ldots, b_{s-1}$.

Proof. This follows immediately from Lemma 4.1.38 and Theorem 4.2.14. □

4.3 Lattice rules

4.3.1 Good lattice points

The discrepancy of a point set is easier to analyze if the point set possesses some structure. There are two popular structures in discrepancy theory, the lattice (or grid) structure considered in this section and the net structure in the sense of Section 4.4.

We have already encountered a point set with an obvious lattice (or grid) structure in dimension s, namely the set of nodes of the s-fold Cartesian product of a midpoint rule. If we start from the midpoint rule for the interval $[0, 1]$ with $m \geq 1$ nodes, then according to (4.14) the corresponding set of s-dimensional nodes is the point set $\mathcal{P}_{m,s}$ consisting of the points

$$\left(\frac{2k_1 - 1}{2m}, \ldots, \frac{2k_s - 1}{2m}\right) \in [0, 1)^s \tag{4.33}$$

with $k_1, \ldots, k_s$ running independently through the integers $1, \ldots, m$. The point set $\mathcal{P}_{m,s}$ contains exactly $N = m^s$ points and is called a *centered regular lattice*. Figure 4.5 illustrates the centered regular lattice with $s = 2$ and $m = 6$.

Intuitively, one may think that the points of $\mathcal{P}_{m,s}$ are very evenly distributed over $[0, 1]^s$, but it turns out that $\mathcal{P}_{m,s}$ is by no means a low-discrepancy point set in the multidimensional case $s \geq 2$. For an ε with $0 < \varepsilon \leq 1/(2m)$, consider the interval $J_\varepsilon = [0, 1 - 1/(2m) + \varepsilon)^s \subseteq [0, 1)^s$ occurring in the definition of the star discrepancy. Since all points of $\mathcal{P}_{m,s}$ are contained in J_ε, we get

$$D_N^*(\mathcal{P}_{m,s}) \geq \left|\frac{A(J_\varepsilon; \mathcal{P}_{m,s})}{N} - \lambda_s(J_\varepsilon)\right| = 1 - \left(1 - \frac{1}{2m} + \varepsilon\right)^s,$$

and letting $\varepsilon \to 0+$ we obtain

$$D_N^*(\mathcal{P}_{m,s}) \geq 1 - \left(1 - \frac{1}{2m}\right)^s.$$

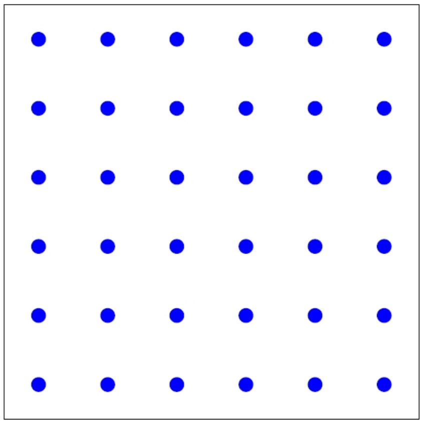

Figure 4.5: The centered regular lattice with $s = 2$ and $m = 6$

Since $0 < 1 - \frac{1}{2m} < 1$, we have $(1 - \frac{1}{2m})^s \leq 1 - \frac{1}{2m}$, and so $D_N^*(\mathcal{P}_{m,s}) \geq \frac{1}{2m} = \frac{1}{2}N^{-1/s}$. Hence for $s \geq 2$, the star discrepancy of $\mathcal{P}_{m,s}$ is asymptotically much larger (in terms of N) than that of a Hammersley point set in dimension s, for instance, with the same number N of points (compare with Theorem 4.2.15).

Therefore we launch another approach in order to arrive at point sets with lattice structure that have a reasonably small (star) discrepancy. Let us start from the Kronecker sequences $(\{n\boldsymbol{\alpha}\})_{n=1}^{\infty}$ in Subsection 4.1.2; here $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_s) \in \mathbb{R}^s$ with $1, \alpha_1, \ldots, \alpha_s$ linearly independent over $\mathbb{Q}$. Now we replace $\boldsymbol{\alpha}$ by a point in $\mathbb{R}^s$ which is in a sense at the other extreme in terms of linear independence over $\mathbb{Q}$, namely a point all of whose coordinates are rational numbers (we may think of such a point also as a discrete approximation of $\boldsymbol{\alpha}$). By putting all its coordinates on the same positive common denominator, such a point can be written in the form $(1/N)\mathbf{g}$ with an integer $N \geq 1$ and $\mathbf{g} \in \mathbb{Z}^s$. The corresponding sequence is then the sequence $(\{(n/N)\mathbf{g}\})_{n=1}^{\infty}$ of fractional parts. It is obvious that this sequence is periodic with period length N, and so we consider only the points in the first period, that is, the points $\{(n/N)\mathbf{g}\}$ with $n = 1, \ldots, N$. We denote this point set by $\mathcal{P}(\mathbf{g}, N)$. Clearly, the lattice point $\mathbf{g}$ matters only modulo N, and so the positive integer N is called the *modulus* of $\mathcal{P}(\mathbf{g}, N)$. We say informally that $\mathbf{g}$ is a *good lattice point* modulo N if the (star) discrepancy of $\mathcal{P}(\mathbf{g}, N)$ is in some sense small. The point sets $\mathcal{P}(\mathbf{g}, N)$ were first introduced by the number theorist Korobov [85] in 1959 and were also proposed independently by Hlawka [64]. Figure 4.6 shows the point set $\mathcal{P}(\mathbf{g}, N)$ with $\mathbf{g} = (1, 21) \in \mathbb{Z}^2$ and $N = 34$.

We want to avoid the trivial case $N = 1$, and so we always assume that $N \geq 2$. The first major issue is to derive a discrepancy bound for the point sets $\mathcal{P}(\mathbf{g}, N)$. This is accomplished by a principle of discrete Fourier analysis for residue class rings of $\mathbb{Z}$. The philosophy of this principle is connected also with the Weyl criterion in

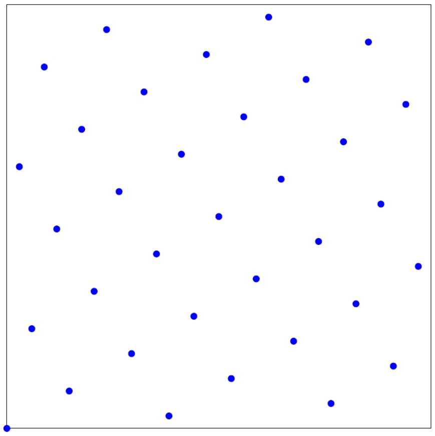

Figure 4.6: A good lattice point set

$\mathbb{R}^s$ (see Theorem 4.1.30).

We need some notation for the formulation and the proof of this principle. For an integer $M \geq 2$, let $C(M) = (-M/2, M/2] \cap \mathbb{Z}$ and put $C^*(M) = C(M) \setminus \{0\}$. Note that $C(M)$ is a complete residue system modulo M which is, as far as possible, symmetric around 0. Furthermore, let $C_s(M)$ be the Cartesian product of s copies of $C(M)$ and put $C_s^*(M) = C_s(M) \setminus \{\mathbf{0}\}$. We set

$$r(h, M) = \begin{cases} M \sin(\pi |h|/M) & \text{for } h \in C^*(M), \\ 1 & \text{for } h = 0. \end{cases}$$

For $\mathbf{h} = (h_1, \ldots, h_s) \in C_s(M)$, we put

$$r(\mathbf{h}, M) = \prod_{i=1}^{s} r(h_i, M).$$

Proposition 4.3.1 *For an integer $M \geq 2$ and for $\mathbf{z}_1, \ldots, \mathbf{z}_N \in \mathbb{Z}^s$, let $\mathcal{P}$ be the point set consisting of the fractional parts $\{M^{-1}\mathbf{z}_1\}, \ldots, \{M^{-1}\mathbf{z}_N\}$. Then*

$$D_N(\mathcal{P}) \leq \frac{s}{M} + \sum_{\mathbf{h} \in C_s^*(M)} \frac{1}{r(\mathbf{h}, M)} \left| \frac{1}{N} \sum_{n=1}^{N} \chi_M(\mathbf{h} \cdot \mathbf{z}_n) \right|,$$

where $\chi_M(z) = e^{2\pi \mathrm{i} z/M}$ for all $z \in \mathbb{Z}$.

Proof. For $\mathbf{k} = (k_1, \ldots, k_s) \in \mathbb{Z}^s$, let $A(\mathbf{k})$ be the number of integers n with $1 \leq n \leq N$ and $\mathbf{z}_n \equiv \mathbf{k} \pmod{M}$, where a congruence between vectors is meant componentwise. Then

$$A(\mathbf{k}) = \sum_{n=1}^{N} \frac{1}{M^s} \sum_{\mathbf{h} \in C_s(M)} \chi_M(\mathbf{h} \cdot (\mathbf{z}_n - \mathbf{k})),$$

since the inner sum has the value M^s if $\mathbf{z}_n \equiv \mathbf{k} \pmod{M}$ and the value 0 otherwise. Therefore

$$A(\mathbf{k}) - \frac{N}{M^s} = \frac{1}{M^s} \sum_{\mathbf{h} \in C_s^*(M)} \chi_M(-\mathbf{h} \cdot \mathbf{k}) \sum_{n=1}^{N} \chi_M(\mathbf{h} \cdot \mathbf{z}_n). \tag{4.34}$$

Now let $J = \prod_{i=1}^{s} [u_i, w_i) \subseteq [0,1)^s$ be an arbitrary interval occurring in the definition of $D_N(\mathcal{P})$. For each $i = 1, \ldots, s$, let $[a_i/M, b_i/M]$ be the largest closed subinterval of $[u_i, w_i)$ with integers $0 \le a_i \le b_i \le M-1$. The case where for some i there is no such subinterval of $[u_i, w_i)$ can be easily dealt with, since then $A(J;\mathcal{P}) = 0$ and $w_i - u_i < 1/M$, hence

$$\Big|\frac{A(J;\mathcal{P})}{N} - \lambda_s(J)\Big| = \lambda_s(J) < \frac{1}{M} \le \frac{s}{M}. \tag{4.35}$$

In the remaining case, the integers $a_1, \ldots, a_s$ and $b_1, \ldots, b_s$ are well defined and we can write

$$\frac{A(J;\mathcal{P})}{N} - \lambda_s(J) = \sum_{\substack{\mathbf{k} \in \mathbb{Z}^s \\ a_i \le k_i \le b_i}} \Big(\frac{A(\mathbf{k})}{N} - \frac{1}{M^s}\Big) + \frac{1}{M^s} \prod_{i=1}^{s} (b_i - a_i + 1) - \lambda_s(J).$$

Now by the choice of the a_i and b_i we obtain

$$\Big|\frac{b_i - a_i + 1}{M} - (w_i - u_i)\Big| \le \frac{1}{M} \qquad \text{for } 1 \le i \le s,$$

and so an application of Lemma 4.2.13 yields

$$\Big|\frac{A(J;\mathcal{P})}{N} - \lambda_s(J)\Big| \le \frac{s}{M} + \Big|\sum_{\substack{\mathbf{k} \in \mathbb{Z}^s \\ a_i \le k_i \le b_i}} \Big(\frac{A(\mathbf{k})}{N} - \frac{1}{M^s}\Big)\Big|. \tag{4.36}$$

Furthermore, (4.34) shows that

$$\Big|\sum_{\substack{\mathbf{k} \in \mathbb{Z}^s \\ a_i \le k_i \le b_i}} \Big(\frac{A(\mathbf{k})}{N} - \frac{1}{M^s}\Big)\Big| \le \frac{1}{M^s} \sum_{\mathbf{h} \in C_s^*(M)} \Big|\sum_{\substack{\mathbf{k} \in \mathbb{Z}^s \\ a_i \le k_i \le b_i}} \chi_M(\mathbf{h} \cdot \mathbf{k})\Big| \Big|\frac{1}{N} \sum_{n=1}^{N} \chi_M(\mathbf{h} \cdot \mathbf{z}_n)\Big|. \tag{4.37}$$

For fixed $\mathbf{h} = (h_1, \ldots, h_s) \in C_s^*(M)$, we can write

$$\Big|\sum_{\substack{\mathbf{k} \in \mathbb{Z}^s \\ a_i \le k_i \le b_i}} \chi_M(\mathbf{h} \cdot \mathbf{k})\Big| = \Big|\sum_{\substack{\mathbf{k} \in \mathbb{Z}^s \\ 0 \le k_i \le b_i - a_i}} \chi_M(\mathbf{h} \cdot \mathbf{k})\Big| = \prod_{i=1}^{s} \Big|\sum_{k_i=0}^{b_i - a_i} \chi_M(h_i k_i)\Big|.$$

If $h_i = 0$, then

$$\Big|\sum_{k_i=0}^{b_i - a_i} \chi_M(h_i k_i)\Big| = b_i - a_i + 1 \le M = \frac{M}{r(h_i, M)}.$$

For $h_i \in C^*(M)$, the summation formula for geometric series and simple trigonometry yield

$$\Big|\sum_{k_i=0}^{b_i-a_i} \chi_M(h_i k_i)\Big| = \Big|\frac{\chi_M(h_i(b_i-a_i+1))-1}{\chi_M(h_i)-1}\Big| = \Big|\frac{\sin(\pi h_i(b_i-a_i+1)/M)}{\sin(\pi h_i/M)}\Big|$$
$$\leq \frac{1}{\sin(\pi|h_i|/M)} = \frac{M}{r(h_i,M)}.$$

Therefore

$$\Big|\sum_{\substack{\mathbf{k}\in\mathbb{Z}^s\\ a_i\leq k_i\leq b_i}} \chi_M(\mathbf{h}\cdot\mathbf{k})\Big| \leq \prod_{i=1}^{s} \frac{M}{r(h_i,M)} = \frac{M^s}{r(\mathbf{h},M)}.$$

We use this in (4.37) and arrive at the inequality

$$\Big|\sum_{\substack{\mathbf{k}\in\mathbb{Z}^s\\ a_i\leq k_i\leq b_i}} \Big(\frac{A(\mathbf{k})}{N} - \frac{1}{M^s}\Big)\Big| \leq \sum_{\mathbf{h}\in C_s^*(M)} \frac{1}{r(\mathbf{h},M)}\Big|\frac{1}{N}\sum_{n=1}^{N}\chi_M(\mathbf{h}\cdot\mathbf{z}_n)\Big|.$$

By inserting this bound into (4.36), we obtain

$$\Big|\frac{A(J;\mathcal{P})}{N} - \lambda_s(J)\Big| \leq \frac{s}{M} + \sum_{\mathbf{h}\in C_s^*(M)} \frac{1}{r(\mathbf{h},M)}\Big|\frac{1}{N}\sum_{n=1}^{N}\chi_M(\mathbf{h}\cdot\mathbf{z}_n)\Big|.$$

In view of (4.35), this inequality holds for all intervals occurring in the definition of $D_N(\mathcal{P})$, and so the desired result follows. □

The second step towards a discrepancy bound for the point sets $\mathcal{P}(\mathbf{g},N)$ is to apply Proposition 4.3.1 to these point sets. As we will see in the proof of Theorem 4.3.3 below, this leads to the quantity introduced in the following definition. It is convenient to put

$$r(\mathbf{h}) = \prod_{i=1}^{s} \max(1,|h_i|) \tag{4.38}$$

for all $\mathbf{h} = (h_1,\ldots,h_s) \in \mathbb{Z}^s$.

Definition 4.3.2 For all $\mathbf{g} \in \mathbb{Z}^s$ and all integers $N \geq 2$, we put

$$R(\mathbf{g},N) = \sum_{\mathbf{h}} \frac{1}{r(\mathbf{h})},$$

where we sum over all $\mathbf{h} \in C_s^*(N)$ with $\mathbf{h}\cdot\mathbf{g} \equiv 0 \pmod{N}$. We use the standard convention that an empty sum is equal to 0.

Theorem 4.3.3 *For all* $\mathbf{g} \in \mathbb{Z}^s$ *and all integers* $N \geq 2$*, the discrepancy of the point set* $\mathcal{P}(\mathbf{g},N)$ *satisfies*

$$D_N(\mathcal{P}(\mathbf{g},N)) \leq \frac{s}{N} + \frac{1}{2}R(\mathbf{g},N).$$

Proof. By Proposition 4.3.1 with $M = N$ and $\mathbf{z}_n = n\mathbf{g}$ for $1 \leq n \leq N$, we obtain

$$D_N(\mathcal{P}(\mathbf{g}, N)) \leq \frac{s}{N} + \sum_{\mathbf{h} \in C_s^*(N)} \frac{1}{r(\mathbf{h}, N)} \Big| \frac{1}{N} \sum_{n=1}^{N} \chi_N(n\mathbf{h} \cdot \mathbf{g}) \Big|.$$

The last sum is easily evaluated: it is equal to N if $\mathbf{h} \cdot \mathbf{g} \equiv 0 \pmod{N}$ and, by the summation formula for geometric series, equal to 0 otherwise. This immediately yields the bound

$$D_N(\mathcal{P}(\mathbf{g}, N)) \leq \frac{s}{N} + \sum_{\substack{\mathbf{h} \in C_s^*(N) \\ \mathbf{h} \cdot \mathbf{g} \equiv 0 \ (\mathrm{mod}\ N)}} \frac{1}{r(\mathbf{h}, N)}.$$

The final result is obtained from $r(\mathbf{h}, N) \geq 2r(\mathbf{h})$ for all $\mathbf{h} \in C_s^*(N)$, which follows in turn from $\sin(\pi u) \geq 2u$ for $0 \leq u \leq \frac{1}{2}$. □

Now that we know a discrepancy bound for the point sets $\mathcal{P}(\mathbf{g}, N)$, we can utilize these point sets in quasi-Monte Carlo integration and we get an error bound by means of the Koksma-Hlawka inequality (see Theorem 4.1.41); note that trivially $D_N^*(\mathcal{P}(\mathbf{g}, N)) \leq D_N(\mathcal{P}(\mathbf{g}, N))$.

There is another approach to error bounds for the point sets $\mathcal{P}(\mathbf{g}, N)$ that exploits the special structure of these point sets. In order to describe this approach, we first embark on a brief excursion into Fourier analysis. The classical setting of Fourier analysis is the one-dimensional case where one considers real-valued periodic functions f on $\mathbb{R}$ with a period that we normalize to be 1. The unit interval $[0, 1]$ is thus a period interval of f. Under a reasonable regularity assumption on f, let us say continuity, we can introduce the *Fourier coefficient* $\hat{f}(h)$ for every $h \in \mathbb{Z}$ by

$$\hat{f}(h) = \int_0^1 f(u) e^{-2\pi \mathrm{i} h u} du.$$

Then we associate with f its formal *Fourier series*

$$\sum_{h \in \mathbb{Z}} \hat{f}(h) e^{2\pi \mathrm{i} h u} \qquad \text{for all } u \in \mathbb{R}.$$

Under an additional hypothesis on f, for instance that the second derivative of f exists and is continuous on $\mathbb{R}$, it can be shown that the Fourier series of f is absolutely convergent and represents f. Hence in this case we get a true identity

$$f(u) = \sum_{h \in \mathbb{Z}} \hat{f}(h) e^{2\pi \mathrm{i} h u} = \lim_{m \to \infty} \sum_{h=-m}^{m} \hat{f}(h) e^{2\pi \mathrm{i} h u} \qquad \text{for all } u \in \mathbb{R}.$$

Remark 4.3.4 If the second derivative f'' of f exists and is continuous on $\mathbb{R}$, then the absolute convergence of the Fourier series of f can be proved easily by using integration by parts. To begin with,

$$\begin{aligned}\hat{f}(h) &= \int_0^1 f(u)e^{-2\pi\mathrm{i}hu}du = \Big[f(u)\frac{e^{-2\pi\mathrm{i}hu}}{-2\pi\mathrm{i}h}\Big]_{u=0}^{1} - \int_0^1 f'(u)\frac{e^{-2\pi\mathrm{i}hu}}{-2\pi\mathrm{i}h}du \\ &= \frac{1}{2\pi\mathrm{i}h}\int_0^1 f'(u)e^{-2\pi\mathrm{i}hu}du\end{aligned}$$

for every nonzero $h \in \mathbb{Z}$. Another integration by parts yields

$$\hat{f}(h) = \frac{1}{(2\pi\mathrm{i}h)^2}\int_0^1 f''(u)e^{-2\pi\mathrm{i}hu}du.$$

Therefore

$$|\hat{f}(h)| \le (2\pi|h|)^{-2} \max_{0\le u\le 1} |f''(u)|\,,$$

and the absolute convergence of the Fourier series of f follows from the convergence of the series $\sum_{h=1}^{\infty} h^{-2}$.

We proceed analogously for an arbitrary dimension $s \ge 1$. Instead of the basic functions $e^{2\pi\mathrm{i}hu}$ (with $h \in \mathbb{Z}$) in one-dimensional Fourier analysis, we use the functions $e^{2\pi\mathrm{i}\mathbf{h}\cdot\mathbf{u}}$ (with $\mathbf{h} \in \mathbb{Z}^s$) in the s-dimensional case. The given real-valued function f on $\mathbb{R}^s$ is now periodic of period 1 in each variable, and so the s-dimensional unit cube $[0,1]^s$ is a period interval of f. Let us assume right away that the function f is sufficiently smooth. For an integer $k \ge 2$, let $C^k(\mathbb{R}^s/\mathbb{Z}^s)$ be the function class consisting of the real-valued periodic functions f on $\mathbb{R}^s$ of period 1 in each variable for which all partial derivatives

$$\frac{\partial^{k_1+\cdots+k_s} f}{\partial u_1^{k_1}\cdots\partial u_s^{k_s}} \qquad \text{with } 0 \le k_i \le k \text{ for } 1 \le i \le s$$

exist and are continuous on $\mathbb{R}^s$. For $f \in C^k(\mathbb{R}^s/\mathbb{Z}^s)$ and $\mathbf{h} \in \mathbb{Z}^s$, we introduce the *Fourier coefficient*

$$\hat{f}(\mathbf{h}) = \int_{[0,1]^s} f(\mathbf{u})e^{-2\pi\mathrm{i}\mathbf{h}\cdot\mathbf{u}}\,d\mathbf{u}.$$

It can be shown by a similar method as in Remark 4.3.4 (but using multidimensional integration by parts when $s \ge 2$) that

$$|\hat{f}(\mathbf{h})| \le c(f)r(\mathbf{h})^{-k} \qquad \text{for all } \mathbf{h} \in \mathbb{Z}^s, \tag{4.39}$$

where the constant $c(f) \ge 0$ depends only on f and where $r(\mathbf{h})$ is as in (4.38). We refer to Zaremba [205], [206, Section 2] for the details. The bound (4.39) readily implies that the *Fourier series*

$$\sum_{\mathbf{h}\in\mathbb{Z}^s} \hat{f}(\mathbf{h})e^{2\pi\mathrm{i}\mathbf{h}\cdot\mathbf{u}}$$

of f is absolutely convergent. It also represents f, and so we arrive at the identity

$$f(\mathbf{u}) = \sum_{\mathbf{h}\in\mathbb{Z}^s} \hat{f}(\mathbf{h})e^{2\pi\mathrm{i}\mathbf{h}\cdot\mathbf{u}} = \lim_{m\to\infty} \sum_{\substack{\mathbf{h}=(h_1,\dots,h_s)\in\mathbb{Z}^s \\ |h_i|\le m}} \hat{f}(\mathbf{h})e^{2\pi\mathrm{i}\mathbf{h}\cdot\mathbf{u}} \qquad \text{for all } \mathbf{u}\in\mathbb{R}^s.$$

Now we examine quasi-Monte Carlo integration for an integrand $f \in C^k(\mathbb{R}^s/\mathbb{Z}^s)$ with a point set $\mathcal{P}(\mathbf{g}, N)$. The quasi-Monte Carlo approximation is

$$\int_{[0,1]^s} f(\mathbf{u})d\mathbf{u} \approx \frac{1}{N}\sum_{n=1}^{N} f\Big(\Big\{\frac{n}{N}\mathbf{g}\Big\}\Big) = \frac{1}{N}\sum_{n=1}^{N} f\Big(\frac{n}{N}\mathbf{g}\Big),$$

where we are allowed to drop the fractional parts since f has period 1 in each variable.

Theorem 4.3.5 *Let $s \ge 1$ be an arbitrary dimension. Let $f \in C^k(\mathbb{R}^s/\mathbb{Z}^s)$ for some integer $k \ge 2$, let $N \ge 2$ be an integer, and let $\mathbf{g} \in \mathbb{Z}^s$. Then*

$$\Big|\int_{[0,1]^s} f(\mathbf{u})d\mathbf{u} - \frac{1}{N}\sum_{n=1}^{N} f\Big(\frac{n}{N}\mathbf{g}\Big)\Big| \le c(f)P_k(\mathbf{g}, N),$$

where $c(f) \ge 0$ is the constant in (4.39) *and where*

$$P_k(\mathbf{g}, N) = \sum_{\mathbf{h}} r(\mathbf{h})^{-k}$$

with the summation running over all nonzero $\mathbf{h} \in \mathbb{Z}^s$ with $\mathbf{h}\cdot\mathbf{g} \equiv 0 \pmod{N}$.

Proof. Since f is represented by its Fourier series and since the value of the integral of f over $[0,1]^s$ is the Fourier coefficient $\hat{f}(\mathbf{0})$, we can write

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N} f\Big(\frac{n}{N}\mathbf{g}\Big) - \int_{[0,1]^s} f(\mathbf{u})d\mathbf{u} &= \frac{1}{N}\sum_{n=1}^{N}\sum_{\mathbf{h}\in\mathbb{Z}^s} \hat{f}(\mathbf{h})e^{2\pi\mathrm{i}(n/N)\mathbf{h}\cdot\mathbf{g}} - \hat{f}(\mathbf{0}) \\ &= \frac{1}{N}\sum_{\mathbf{h}\in\mathbb{Z}^s} \hat{f}(\mathbf{h})\sum_{n=1}^{N} e^{2\pi\mathrm{i}(n/N)\mathbf{h}\cdot\mathbf{g}} - \hat{f}(\mathbf{0}) \\ &= \frac{1}{N}\sum_{\mathbf{h}\in\mathbb{Z}^s,\,\mathbf{h}\neq\mathbf{0}} \hat{f}(\mathbf{h})\sum_{n=1}^{N} e^{2\pi\mathrm{i}(n/N)\mathbf{h}\cdot\mathbf{g}}.\end{aligned}$$

Now using the bound (4.39), we obtain

$$\Big|\int_{[0,1]^s} f(\mathbf{u})d\mathbf{u} - \frac{1}{N}\sum_{n=1}^{N} f\Big(\frac{n}{N}\mathbf{g}\Big)\Big| \le \frac{c(f)}{N}\sum_{\mathbf{h}\in\mathbb{Z}^s,\,\mathbf{h}\neq\mathbf{0}} r(\mathbf{h})^{-k}\Big|\sum_{n=1}^{N}\chi_N(n\mathbf{h}\cdot\mathbf{g})\Big|$$

with the notation in Proposition 4.3.1. By an observation in the proof of Theorem 4.3.3, the last sum is equal to N if $\mathbf{h}\cdot\mathbf{g} \equiv 0 \pmod{N}$ and equal to 0 otherwise.

This leads immediately to the desired result. □

As it stands, Theorem 4.3.5 holds only for periodic integrands, but there are periodization techniques by which it can be extended to nonperiodic integrands. The simplest idea in the one-dimensional case is to take a function f on $[0, 1]$ and replace it by the function f_1 given by

$$f_1(u) = \frac{1}{2}(f(u) + f(1-u)) \qquad \text{for } 0 \le u \le 1.$$

Since $f_1(0) = f_1(1)$, the function f_1 can be extended periodically to $\mathbb{R}$ with period 1. Furthermore, if f is Riemann-integrable on $[0, 1]$, then so is f_1 and

$$\int_0^1 f_1(u)du = \int_0^1 f(u)du.$$

Therefore the numerical integration problem for f is the same as that for f_1, but f_1 is periodic. In this sense, periodicity is not a serious restriction in numerical integration. More sophisticated periodization techniques, also for the multidimensional case, are covered in Sloan and Joe [189, Section 2.12] and Zaremba [206, Section 3]. Another approach to using the method of good lattice points for nonperiodic integrands is based on modified vertex weights (see [144] and [189, Chapter 8]).

We have now two quantities governing the error in quasi-Monte Carlo integration with a point set $\mathcal{P}(\mathbf{g}, N)$: the number $R(\mathbf{g}, N)$ furnishing the discrepancy bound in Theorem 4.3.3 and the number $P_k(\mathbf{g}, N)$ yielding the error bound for integrands in $C^k(\mathbb{R}^s/\mathbb{Z}^s)$ according to Theorem 4.3.5. In order to obtain a small value of $R(\mathbf{g}, N)$, and thus a small discrepancy bound, we have to choose the lattice point $\mathbf{g} \in \mathbb{Z}^s$ in such a way that all summands $r(\mathbf{h})^{-1}$ in the definition of $R(\mathbf{g}, N)$ (see Definition 4.3.2) are small, or equivalently, that the values of $r(\mathbf{h})$ are large for all $\mathbf{h} \in C_s^*(N)$ with $\mathbf{h} \cdot \mathbf{g} \equiv 0 \pmod{N}$. A similar strategy applies to $P_k(\mathbf{g}, N)$ in view of its definition in Theorem 4.3.5: in order to make $P_k(\mathbf{g}, N)$ small, choose $\mathbf{g} \in \mathbb{Z}^s$ in such a way that $r(\mathbf{h})$ is large for all nonzero $\mathbf{h} \in \mathbb{Z}^s$ with $\mathbf{h} \cdot \mathbf{g} \equiv 0 \pmod{N}$. This analogy between these quantities makes it plausible that $P_k(\mathbf{g}, N)$ can be bounded in terms of $R(\mathbf{g}, N)$. The proof of the following inequality from [134, Theorem 5.5] is quite technical, and so we state this result without proof.

Proposition 4.3.6 *Let $k \ge 2$, $s \ge 1$, and $N \ge 2$ be integers, and let $\mathbf{g} \in \mathbb{Z}^s$ be such that each coordinate of $\mathbf{g}$ is coprime to N. Then*

$$P_k(\mathbf{g}, N) \le R(\mathbf{g}, N)^k + c(k, s)N^{-k}$$

with a constant $c(k, s) > 0$ depending only on k and s.

There is a crucial issue remaining, namely how small, for given integers $s \ge 1$ and $N \ge 2$, we can make $R(\mathbf{g}, N)$ by a suitable choice of the lattice point $\mathbf{g} \in \mathbb{Z}^s$. Such an advantageous choice of $\mathbf{g}$ corresponds to what we earlier called a good lattice point modulo N. The trouble is that, except for easy low-dimensional cases (see Example 4.3.15 below), no explicit constructions of good lattice points modulo N are available. Finding such explicit constructions in arbitrary dimensions is indeed *the* outstanding open problem in the theory of good lattice points, and it seems to be a hard nut to crack.

One way around this difficulty is to randomize the problem in some sense, that is, we investigate the average quality of the lattice points $\mathbf{g} \in \mathbb{Z}^s$ (in terms of $R(\mathbf{g}, N)$) for fixed s and N. Thus, instead of trying to reach for the absolute minimum of $R(\mathbf{g}, N)$ for fixed s and N, we are less ambitious and settle for the average value of $R(\mathbf{g}, N)$. To our great relief, this average value is reasonably small. The case $s = 1$ is trivial since then $\mathbf{g}$ has only one coordinate which we take to be 1 (or any integer coprime to N); then $R(\mathbf{g}, N) = 0$ by the convention in Definition 4.3.2. The analysis of the average value of $R(\mathbf{g}, N)$ is considerably easier if N is a prime number. Since $\mathbf{g}$ matters only modulo N, it suffices to average $R(\mathbf{g}, N)$ over $\mathbf{g} \in C_s(N)$.

Theorem 4.3.7 *Let $s \geq 2$ be a dimension and let N be a prime number. Then*

$$M_s(N) := \frac{1}{N^s} \sum_{\mathbf{g} \in C_s(N)} R(\mathbf{g}, N) < \frac{1}{N}(2 \log N + 2)^s.$$

Proof. By the definition of $R(\mathbf{g}, N)$ in Definition 4.3.2, we get

$$M_s(N) = \frac{1}{N^s} \sum_{\mathbf{g} \in C_s(N)} \sum_{\substack{\mathbf{h} \in C_s^*(N) \\ \mathbf{h} \cdot \mathbf{g} \equiv 0 \ (\mathrm{mod}\ N)}} \frac{1}{r(\mathbf{h})} = \frac{1}{N^s} \sum_{\mathbf{h} \in C_s^*(N)} \frac{S(\mathbf{h})}{r(\mathbf{h})},$$

where $S(\mathbf{h})$ is the number of $\mathbf{g} \in C_s(N)$ with $\mathbf{h} \cdot \mathbf{g} \equiv 0 \pmod{N}$. If we write $\mathbf{h} = (h_1, \dots, h_s)$ and $\mathbf{g} = (g_1, \dots, g_s)$, then the last condition means that

$$h_1 g_1 + \cdots + h_s g_s \equiv 0 \pmod{N}. \tag{4.40}$$

If $\mathbf{h} \in C_s^*(N)$, then $h_i \not\equiv 0 \pmod{N}$ for some i with $1 \leq i \leq s$. Thus, for every choice of $g_1, \dots, g_{i-1}, g_{i+1}, \dots, g_s \in C(N)$, the value of $g_i \in C(N)$ is uniquely determined by (4.40) since N is a prime number. Therefore $S(\mathbf{h}) = N^{s-1}$ for all $\mathbf{h} \in C_s^*(N)$ and

$$M_s(N) = \frac{1}{N} \sum_{\mathbf{h} \in C_s^*(N)} \frac{1}{r(\mathbf{h})} < \frac{1}{N} \sum_{\mathbf{h} \in C_s(N)} \frac{1}{r(\mathbf{h})}.$$

By invoking (4.38), we get

$$\sum_{\mathbf{h} \in C_s(N)} \frac{1}{r(\mathbf{h})} = \Big(\sum_{h \in C(N)} \frac{1}{\max(1, |h|)} \Big)^s.$$

For $N \geq 3$, we use the standard method of comparing sums and integrals to obtain

$$\begin{aligned}\sum_{h \in C(N)} \frac{1}{\max(1, |h|)} &= 3 + 2 \sum_{h=2}^{(N-1)/2} \frac{1}{h} \leq 3 + 2 \int_1^{(N-1)/2} \frac{du}{u} \\ &= 3 + 2 \log \frac{N-1}{2} < 2 \log N + 2.\end{aligned}$$

This bound holds trivially for $N = 2$. A combination of the inequalities and identities above yields the theorem. □

Corollary 4.3.8 *For every dimension $s \geq 2$ and every prime number N, there exists a lattice point $\mathbf{g} \in \mathbb{Z}^s$ with*

$$R(\mathbf{g}, N) < \frac{1}{N}(2 \log N + 2)^s.$$

Proof. This follows immediately from Theorem 4.3.7. □

Corollary 4.3.9 *For every dimension $s \geq 2$ and every prime number N, there exists a lattice point $\mathbf{g} \in \mathbb{Z}^s$ for which the discrepancy of the point set $\mathcal{P}(\mathbf{g}, N)$ satisfies*

$$D_N(\mathcal{P}(\mathbf{g}, N)) < \frac{1}{2N}(2 \log N + 2)^s + \frac{s}{N}.$$

Proof. This follows immediately from Theorem 4.3.3 and Corollary 4.3.8. □

Corollary 4.3.10 *For every dimension $s \geq 2$ and every prime number N, there exists a lattice point $\mathbf{g} \in \mathbb{Z}^s$ for which*

$$P_k(\mathbf{g}, N) \leq c_1(k, s) N^{-k} (\log N)^{ks} \qquad \textit{for all integers } k \geq 2,$$

where the constant $c_1(k, s) > 0$ depends only on k and s.

Proof. By averaging in Theorem 4.3.7 not over all $\mathbf{g} \in C_s(N)$, but only over the $(N-1)^s$ lattice points $\mathbf{g} \in C_s(N)$ with all coordinates nonzero, we get such a $\mathbf{g}$ with $R(\mathbf{g}, N) = O(N^{-1}(\log N)^s)$, where the implied constant depends only on s. The rest follows from Proposition 4.3.6. □

Let us now ponder the practical implications of the last two corollaries. Corollary 4.3.9 yields point sets consisting of N points in $[0, 1)^s$, with a prime number N, for which the discrepancy, and therefore also the star discrepancy, is at most of the order of magnitude $N^{-1}(\log N)^s$. These point sets are not necessarily low-discrepancy point sets in the technical sense of (4.23), but their star discrepancy is off by at most a factor $\log N$ from the bound in (4.23). These point sets are therefore serviceable for quasi-Monte Carlo integration and they have the sympathetic feature that their points are very easy to compute once the lattice point $\mathbf{g}$ is known.

Corollary 4.3.10 shows that there are good lattice points $\mathbf{g} \in \mathbb{Z}^s$ modulo a prime number N for which the quasi-Monte Carlo method in Theorem 4.3.5 has a rate of convergence $N^{-k}(\log N)^{ks}$ for integrands $f \in C^k(\mathbb{R}^s/\mathbb{Z}^s)$. The exponent ks of $\log N$ can be improved to $k(s-1)$ by a refined argument due to Bakhvalov [7]. The rate of convergence N^{-k} (up to logarithmic factors) for $f \in C^k(\mathbb{R}^s/\mathbb{Z}^s)$ is really gratifying since it says that the convergence is faster for smoother integrands. This phenomenon is well known in the one-dimensional case (see [35, Chapter 2]). The method of good lattice points extends this phenomenon to the multidimensional case in an elegant fashion. All this is in sharp contrast to the Monte Carlo method where the rate of convergence is $N^{-1/2}$ no matter how smooth the integrand is.

Remark 4.3.11 We restricted the discussion of existence theorems for good lattice points to prime moduli N since this case is much easier to handle. However, existence theorems for good lattice points of the same quality in terms of the orders of magnitude are available also for composite moduli N. This generalization was achieved by Niederreiter [126]. The basic step is to consider, for every integer $N \geq 2$, the set

$$G_s(N) = \{\mathbf{g} = (g_1, \ldots, g_s) \in C_s(N) : \gcd(g_i, N) = 1 \text{ for } 1 \leq i \leq s\} \tag{4.41}$$

of lattice points. The set $G_s(N)$ has exactly $\phi(N)^s$ elements, where ϕ is Euler's totient function. Then the analog of Theorem 4.3.7 says that for all integers $s \geq 2$ and $N \geq 2$, the corresponding average value $A_s(N)$ satisfies

$$A_s(N) := \frac{1}{\phi(N)^s} \sum_{\mathbf{g} \in G_s(N)} R(\mathbf{g}, N) = O(N^{-1}(\log N)^s)$$

with an implied constant depending only on s. An even sharper result providing an asymptotic expansion for $A_s(N)$ can be found in [134, Theorem 5.10]. This yields right away analogs of Corollaries 4.3.8 and 4.3.9 for all integers $N \geq 2$. Since the condition on $\mathbf{g}$ in Proposition 4.3.6 is satisfied for all $\mathbf{g} \in G_s(N)$, we also get an analog of Corollary 4.3.10, namely that for all integers $s \geq 2$ and $N \geq 2$ there exists a lattice point $\mathbf{g} \in G_s(N)$ for which

$$P_k(\mathbf{g}, N) = O(N^{-k}(\log N)^{ks}) \qquad \text{for all integers } k \geq 2,$$

where the implied constant depends only on k and s. For many values of N, the exponent ks of $\log N$ can be improved to $k(s-1)+1$ or even $k(s-1)$ (see [135]).

All existence theorems for good lattice points presented so far are nonconstructive, since they are based on the argument that there must exist a lattice point $\mathbf{g}_0$ for which $R(\mathbf{g}_0, N)$ is at least as small as the average of $R(\mathbf{g}, N)$ over a set of candidate lattice points $\mathbf{g}$. But no information is given in the proofs about how to obtain such a good lattice point $\mathbf{g}_0$. For a long time, good lattice points were produced by brute-force computer search, and this approach is feasible as long as the dimension s and the modulus N are not too large. Note that if we search for an s-dimensional good lattice point modulo N in the set $C_s(N)$, then *a priori* we have to consider N^s candidates.

There is a more constructive strategy that builds an s-dimensional good lattice point coordinate by coordinate, starting with the first coordinate and ending with the sth coordinate. The corresponding algorithm is known in the literature as the *CBC* (for component-by-component) *algorithm*. It is a type of greedy algorithm which proceeds by finding, in a certain sense, local minima of the quantity $R(\mathbf{g}, N)$. It is convenient to put

$$G(N) = \{g \in C(N) : \gcd(g, N) = 1\}.$$

Algorithm 4.3.12 (CBC Algorithm) Let the integers $s \geq 2$ and $N \geq 2$ be given.
Step 1: choose $g_1 = 1$.
Step 2: Suppose that for some dimension d with $1 \leq d \leq s-1$, the coordinates $g_1, \ldots, g_d \in G(N)$ have already been constructed. Then find an integer $g_{d+1} \in G(N)$ that minimizes $R((g_1, \ldots, g_d, b), N)$ as a function of $b \in G(N)$. This recursive procedure stops once the coordinate g_s has been obtained.

The final output of the CBC algorithm is a lattice point $\mathbf{g} = (g_1, \ldots, g_s) \in G(N)^s = G_s(N)$, in the notation of (4.41). In the course of the CBC algorithm we compute $(s-1)\phi(N)$ values of $R(\cdot, N)$, whereas in a brute-force search over the whole set $G_s(N)$ we compute $\phi(N)^s$ values of $R(\cdot, N)$. Therefore the CBC algorithm is much more efficient than brute-force search.

What can we say about the lattice point $\mathbf{g} \in G_s(N)$ produced by the CBC algorithm? The following result provides an answer for the simplest case where N is a prime number.

Theorem 4.3.13 *Let $s \geq 2$ be a given dimension and let N be a prime number. Then the lattice point $\mathbf{g} \in G_s(N)$ produced by the CBC algorithm satisfies*

$$R(\mathbf{g}, N) < \frac{1}{N-1}(2\log N + 2)^s.$$

Proof. We prove by induction on $d = 1, \ldots, s$ that

$$R((g_1, \ldots, g_d), N) < \frac{1}{N-1}(2\log N + 2)^d. \tag{4.42}$$

This is trivial for $d = 1$ since $R(g_1, N) = R(1, N) = 0$. Suppose that (4.42) has been shown for some d with $1 \leq d \leq s - 1$. Note that $G(N) = C^*(N)$ since N is a prime number. Therefore by Step 2 in Algorithm 4.3.12 with $\mathbf{g}_d = (g_1, \dots, g_d)$ and $\mathbf{h} \in C_d(N)$,

$$\begin{aligned} R((\mathbf{g}_d, g_{d+1}), N) &\leq \frac{1}{N-1} \sum_{b \in C^*(N)} R((\mathbf{g}_d, b), N) \\ &= \frac{1}{N-1} \sum_{b \in C^*(N)} \sum_{\substack{(\mathbf{h},k) \in C^*_{d+1}(N) \\ (\mathbf{h},k)\cdot(\mathbf{g}_d, b) \equiv 0 \ (\mathrm{mod}\ N)}} \frac{1}{r(\mathbf{h}) \max(1, |k|)} \\ &= \frac{1}{N-1} \sum_{(\mathbf{h},k) \in C^*_{d+1}(N)} \frac{1}{r(\mathbf{h}) \max(1, |k|)} \sum_{\substack{b \in C^*(N) \\ (\mathbf{h},k)\cdot(\mathbf{g}_d, b) \equiv 0 \ (\mathrm{mod}\ N)}} 1. \end{aligned}$$

The contribution of the terms with $k = 0$ in the last double sum is equal to $(N - 1)R(\mathbf{g}_d, N)$. By splitting off these terms, we get with $\mathbf{g}_{d+1} = (g_1, \dots, g_d, g_{d+1})$,

$$R(\mathbf{g}_{d+1}, N) \leq R(\mathbf{g}_d, N) + \frac{1}{N-1} \sum_{\mathbf{h} \in C_d(N)} \frac{1}{r(\mathbf{h})} \sum_{k \in C^*(N)} \frac{1}{|k|} \sum_{\substack{b \in C^*(N) \\ (\mathbf{h},k)\cdot(\mathbf{g}_d, b) \equiv 0 \ (\mathrm{mod}\ N)}} 1.$$

For fixed $\mathbf{h} \in C_d(N)$ and $k \in C^*(N)$, the congruence $(\mathbf{h}, k) \cdot (\mathbf{g}_d, b) = \mathbf{h} \cdot \mathbf{g}_d + kb \equiv 0 \pmod{N}$ has at most one solution $b \in C^*(N)$ since N is a prime number. Hence using the induction hypothesis (4.42) and bounds in the proof of Theorem 4.3.7, we obtain

$$\begin{aligned} R(\mathbf{g}_{d+1}, N) &\leq R(\mathbf{g}_d, N) + \frac{1}{N-1} \sum_{\mathbf{h} \in C_d(N)} \frac{1}{r(\mathbf{h})} \sum_{k \in C^*(N)} \frac{1}{|k|} \\ &< \frac{1}{N-1}(2 \log N + 2)^d + \frac{1}{N-1}(2 \log N + 2)^d (2 \log N + 1) \\ &= \frac{1}{N-1}(2 \log N + 2)^{d+1}. \end{aligned}$$

This completes the proof of (4.42) by induction, and putting $d = s$ in (4.42) yields the result of the theorem. □

It is remarkable that Theorem 4.3.13 is practically of the same quality as the nonconstructive existence result in Corollary 4.3.8. Clearly, Corollary 4.3.9 with $\frac{1}{2N}$ replaced by $\frac{1}{2(N-1)}$ and Corollary 4.3.10 hold also for the lattice point $\mathbf{g} \in G_s(N)$ in Theorem 4.3.13. An extension of Theorem 4.3.13 to composite moduli N was achieved by Sinescu and Joe [186] and it yields again the order of magnitude $N^{-1}(\log N)^s$ for the quantity $R(\mathbf{g}, N)$ with $\mathbf{g} \in G_s(N)$ being the lattice point produced by the CBC algorithm.

Now we come to another fruitful idea in the search for good lattice points, namely to reduce the size of the search space by restricting the form of the lattice points.

The most popular special form is

$$\mathbf{g}(a) := (1, a, a^2, \ldots, a^{s-1}) \in \mathbb{Z}^s \tag{4.43}$$

which is called the *Korobov form* because it was proposed by Korobov [86]. The lattice point $\mathbf{g}(a)$ depends also on s, but for the sake of simplicity we suppress this dependence in the notation; the dimension s will always be clear from the context. Since $\mathbf{g}(a)$ matters only modulo a given integer $N \geq 2$, we can confine the integer a to a complete residue system modulo N, say $a \in Z_N = \{0, 1, \ldots, N-1\}$ or $a \in C(N)$. Hence there are only N candidates $\mathbf{g}(a)$ in the search space, as opposed to N^s candidates when the lattice points run through the set $C_s(N)$. The following result, which is an analog of Theorem 4.3.7, guarantees that the strategy of limiting the search to lattice points of Korobov form is successful, at least in the case of prime moduli.

Theorem 4.3.14 *Let $s \geq 2$ be a dimension and let N be a prime number. Then*

$$K_s(N) := \frac{1}{N} \sum_{a=0}^{N-1} R(\mathbf{g}(a), N) < \frac{s-1}{N}(2 \log N + 2)^s.$$

Proof. By inserting the definition of $R(\mathbf{g}(a), N)$ into the expression for $K_s(N)$, we obtain as in the proof of Theorem 4.3.7 that

$$K_s(N) = \frac{1}{N} \sum_{\mathbf{h} \in C_s^*(N)} \frac{T(\mathbf{h})}{r(\mathbf{h})},$$

where $T(\mathbf{h})$ is the number of $a \in Z_N$ with $\mathbf{h} \cdot \mathbf{g}(a) \equiv 0 \pmod{N}$. If we write $\mathbf{h} = (h_1, \ldots, h_s)$, then the last condition means that

$$\mathbf{h} \cdot \mathbf{g}(a) = h_1 + h_2 a + h_3 a^2 + \cdots + h_s a^{s-1} \equiv 0 \pmod{N}.$$

For fixed $\mathbf{h} \in C_s^*(N)$, this is a polynomial congruence modulo N in the unknown a with a nonzero polynomial of degree at most $s-1$. Since N is a prime number, it follows that $T(\mathbf{h}) \leq s-1$ (apply Theorem 1.4.27 to the finite field $\mathbb{F}_N$). We infer that

$$K_s(N) \leq \frac{s-1}{N} \sum_{\mathbf{h} \in C_s^*(N)} \frac{1}{r(\mathbf{h})} < \frac{s-1}{N}(2 \log N + 2)^s$$

by a bound in the proof of Theorem 4.3.7. □

In terms of N, the upper bound in Theorem 4.3.14 has the same order of magnitude $N^{-1}(\log N)^s$ as that in Theorem 4.3.7. Therefore Theorem 4.3.14 has consequences like Corollaries 4.3.8, 4.3.9, and 4.3.10 for lattice points of Korobov form, with upper bounds of the same order of magnitude in N as in those corollaries.

Example 4.3.15 Except for the trivial one-dimensional case, there is only the two-dimensional case in which general explicit constructions of good lattice points are known. The nicest such construction uses Fibonacci numbers. Recall that the sequence $F_1, F_2, \ldots$ of Fibonacci numbers is the sequence of positive integers defined recursively by $F_1 = F_2 = 1$ and $F_{k+2} = F_{k+1} + F_k$ for $k \geq 1$. Thus $F_3 = 2$, $F_4 = 3$, $F_5 = 5$, $F_6 = 8$, and so on. The Fibonacci sequence is closely connected with the irrational number $\alpha = (\sqrt{5} - 1)/2$ considered in Example 4.2.7. We showed in that example that α has the periodic continued fraction expansion

$$\alpha = [0; 1, 1, 1, \ldots].$$

Then by the recursions stated prior to Lemma 4.2.1, the numerators p_k and the denominators q_k of the convergents p_k/q_k to α are given by $p_k = F_k$ and $q_k = F_{k+1}$ for all $k \geq 1$. This suggests the following construction of two-dimensional good lattice points. As a modulus we take $N = F_m$ for some integer $m \geq 3$. Note that there are composite numbers among the Fibonacci numbers, for example $F_6 = 8$ and $F_8 = 21$, and so the restriction to prime moduli as in most of our discussion of good lattice points is not needed here. The lattice point corresponding to the modulus $N = F_m$ is $\mathbf{g} = (1, F_{m-1})$. This lattice point is of Korobov form. We will show that $\mathbf{g}$ is a good lattice point modulo N in the strong sense that $\mathcal{P}(\mathbf{g}, N)$ is actually a two-dimensional low-discrepancy point set. First we consider the sequence $\mathcal{S} = (\{nF_{m-1}/F_m\})_{n=0}^{\infty}$. We make the crucial observation that in the proof of the discrepancy bound in Theorem 4.2.5, the fact that α is irrational is not used explicitly, but only the properties of the convergents to α are relevant. This entails that the argument in the proof of Theorem 4.2.5 applies also to the first M terms of the sequence $\mathcal{S}$ as long as $M \leq N = F_m$. Therefore the consequence of Theorem 4.2.5 stated in Theorem 4.2.6 applies as well for this range of M. Since in our case $K = 1$ in Theorem 4.2.6, we obtain

$$D_M(\mathcal{S}) < \frac{2}{\log 2} M^{-1} \log(M+1) \qquad \text{for } 1 \leq M \leq N.$$

Next we recall that $\mathcal{P}(\mathbf{g}, N)$ consists of the points $(\{n/F_m\}, \{nF_{m-1}/F_m\}) \in [0,1)^2$ with $n = 1, \ldots, N = F_m$. Note that for $n = F_m$ we get $(\{n/F_m\}, \{nF_{m-1}/F_m\}) = (0,0)$, and so $\mathcal{P}(\mathbf{g}, N)$ can be described also as the point set consisting of the points

$$\left(\frac{n}{F_m}, \left\{\frac{nF_{m-1}}{F_m}\right\}\right) \in [0,1)^2 \qquad \text{for } n = 0, 1, \ldots, N-1 = F_m - 1.$$

Therefore we can apply Lemma 4.1.38, and using the trivial fact that $D_M^*(\mathcal{S}) \leq D_M(\mathcal{S})$ for all $M \geq 1$ we get

$$D_N^*(\mathcal{P}(\mathbf{g}, N)) < \frac{2}{\log 2} \cdot \frac{\log(N+1)}{N} + \frac{1}{N}.$$

Thus, $\mathcal{P}(\mathbf{g}, N)$ is indeed a two-dimensional low-discrepancy point set in the sense of (4.23). This is excellent news since it yields an improvement on the quality of lattice points promised by the existence result in Corollary 4.3.9 and by the CBC algorithm in Theorem 4.3.13 for $s = 2$. A detailed analysis of the special two-dimensional point sets $\mathcal{P}(\mathbf{g}, N)$ built on Fibonacci numbers can be found in Zaremba [204]. The case where $N = F_9 = 34$ is illustrated in Figure 4.6. Unfortunately, nobody has managed to extend this elegant construction to higher dimensions in such a way that it at least matches the existence result in Corollary 4.3.9.

4.3.2 General lattice rules

The issue here is to generalize the method of good lattice points so as to bring out its salient features. We commence by regarding this method from a group-theoretic perspective. For a given dimension $s \geq 1$, the Euclidean space $\mathbb{R}^s$ is an abelian group under addition (of real numbers for $s = 1$ and of vectors for $s \geq 2$). Since $\mathbb{Z}^s$ is a subgroup of $\mathbb{R}^s$, we can form the factor group $\mathbb{R}^s/\mathbb{Z}^s$ which is sometimes called the s-dimensional *torus group*. For $s = 1$ the set $\mathbb{R}/\mathbb{Z}$ is geometrically similar to a circle since the endpoints 0 and 1 of the unit interval $[0, 1]$ belong to the same coset $0+\mathbb{Z} \in \mathbb{R}/\mathbb{Z}$ and can therefore be identified. Similarly for $s = 2$, we can think of the set $\mathbb{R}^2/\mathbb{Z}^2$ as being obtained by identifying opposite sides of the unit square $[0, 1]^2$, and then we arrive at the geometric shape of a doughnut (or a torus in the technical jargon). For $s \geq 3$ we get higher-dimensional tori (not to be confused with torii which, as everybody knows, is a gateway of a Shintō shrine) .

Now let us consider a point set $\mathcal{P}(\mathbf{g}, N)$ in Subsection 4.3.1 with a dimension $s \geq 1$, a lattice point $\mathbf{g} \in \mathbb{Z}^s$, and a modulus $N \geq 2$. By definition, the points of $\mathcal{P}(\mathbf{g}, N)$ are the fractional parts $\mathbf{x}_n := \{(n/N)\mathbf{g}\}$ with $n = 1, \ldots, N$. The corresponding cosets in $\mathbb{R}^s/\mathbb{Z}^s$ are given by

$$\mathbf{x}_n + \mathbb{Z}^s = \left\{\frac{n}{N}\mathbf{g}\right\} + \mathbb{Z}^s = \frac{n}{N}\mathbf{g} + \mathbb{Z}^s = n\Big(\frac{1}{N}\mathbf{g} + \mathbb{Z}^s\Big) \qquad \text{for } n = 1, \ldots, N.$$

For $n = N$ we have $\mathbf{x}_N + \mathbb{Z}^s = \mathbf{g} + \mathbb{Z}^s = \mathbf{0} + \mathbb{Z}^s$, the identity element of the group $\mathbb{R}^s/\mathbb{Z}^s$. Therefore the cosets $\mathbf{x}_n+\mathbb{Z}^s$ for $n = 1, \ldots, N$ form the finite cyclic subgroup of $\mathbb{R}^s/\mathbb{Z}^s$ generated by $(1/N)\mathbf{g} + \mathbb{Z}^s$.

If the point set $\mathcal{P}(\mathbf{g}, N)$ is viewed in this way, then the following generalization is obvious. Let $L/\mathbb{Z}^s$ be *any* finite subgroup of $\mathbb{R}^s/\mathbb{Z}^s$ and let $\mathbf{y}_n+\mathbb{Z}^s$ with $\mathbf{y}_n \in [0, 1)^s$ for $n = 1, \ldots, N$ be the distinct cosets making up the group $L/\mathbb{Z}^s$. The point set consisting of the points $\mathbf{y}_1, \ldots, \mathbf{y}_N$ is called a *lattice point set* and the corresponding quasi-Monte Carlo approximation

$$\int_{[0,1]^s} f(\mathbf{u})d\mathbf{u} \approx \frac{1}{N}\sum_{n=1}^{N} f(\mathbf{y}_n) \tag{4.44}$$

is called a *lattice rule.*

The name “lattice rule” stems from a geometric interpretation of the group-theoretic approach above. If we envisage the union $L = \cup_{n=1}^{N}(\mathbf{y}_n + \mathbb{Z}^s)$ of cosets as a subset of $\mathbb{R}^s$, then L is an s-dimensional lattice. Here by an s-dimensional *lattice* we mean a discrete additive subgroup of $\mathbb{R}^s$ that is not contained in any proper linear subspace of $\mathbb{R}^s$. Equivalently, an s-dimensional lattice is obtained by taking s linearly independent vectors $\mathbf{b}_1, \ldots, \mathbf{b}_s \in \mathbb{R}^s$ (hence a basis of the vector space $\mathbb{R}^s$) and forming the set

$$L = \Big\{ \sum_{i=1}^{s} k_i \mathbf{b}_i : k_i \in \mathbb{Z} \ \text{ for } 1 \le i \le s \Big\} \tag{4.45}$$

of all linear combinations of $\mathbf{b}_1, \ldots, \mathbf{b}_s$ with coefficients that are integers. The lattices corresponding to lattice rules must have an additional property stipulated in the following definition.

Definition 4.3.16 An s-dimensional lattice is called an s-dimensional *integration lattice* if it contains $\mathbb{Z}^s$ as a subset.

Instead of a finite subgroup $L/\mathbb{Z}^s$ of $\mathbb{R}^s/\mathbb{Z}^s$, we can then take an s-dimensional integration lattice L as the starting point. The intersection $L \cap [0,1)^s$ is a finite set since L is discrete, and this finite set of points in $[0,1)^s$ forms again a lattice point set.

The cornerstone for the analysis of the discrepancy of general lattice point sets is again Proposition 4.3.1, as in the special case of the point sets $\mathcal{P}(\mathbf{g}, N)$. This leads naturally to the following concept.

Definition 4.3.17 The *dual lattice* $L^{\perp}$ of the s-dimensional integration lattice L is defined by

$$L^{\perp} = \{\mathbf{h} \in \mathbb{Z}^s : \mathbf{h} \cdot \mathbf{y} \in \mathbb{Z} \ \text{ for all } \mathbf{y} \in L\}.$$

Example 4.3.18 Let us consider the special case of the integration lattice L corresponding to a point set $\mathcal{P}(\mathbf{g}, N)$. We know that

$$L = \bigcup_{n=1}^{N} \Big(\frac{n}{N} \mathbf{g} + \mathbb{Z}^s \Big).$$

Thus, the elements $\mathbf{y} \in L$ are exactly given by $\mathbf{y} = (n/N)\mathbf{g} + \mathbf{k}$ for some $n = 1, \ldots, N$ and some $\mathbf{k} \in \mathbb{Z}^s$. For $\mathbf{h} \in \mathbb{Z}^s$ we therefore obtain $\mathbf{h} \in L^{\perp}$ if and only if $(n/N)\mathbf{h} \cdot \mathbf{g} + \mathbf{h} \cdot \mathbf{k} \in \mathbb{Z}$ for all $n = 1, \ldots, N$ and all $\mathbf{k} \in \mathbb{Z}^s$. But $\mathbf{h} \cdot \mathbf{k}$ is automatically an integer, and so the last condition is equivalent to $(n/N)\mathbf{h} \cdot \mathbf{g} \in \mathbb{Z}$ for $n = 1, \ldots, N$. It suffices to require this for $n = 1$, hence the condition says that $(1/N)\mathbf{h} \cdot \mathbf{g} \in \mathbb{Z}$, and therefore

$$L^{\perp} = \{\mathbf{h} \in \mathbb{Z}^s : \mathbf{h} \cdot \mathbf{g} \equiv 0 \ (\text{mod } N)\}.$$

It is no accident that we have already seen this condition in the analysis of the point sets $\mathcal{P}(\mathbf{g}, N)$, for instance in Definition 4.3.2 and Theorem 4.3.5. The dual lattice plays a crucial role in the analysis of general lattice point sets as well.

Lemma 4.3.19 *Let $\mathbf{y}_1, \dots, \mathbf{y}_N \in [0,1)^s$ be the points of the lattice point set corresponding to the s-dimensional integration lattice L, or equivalently to the finite subgroup $L/\mathbb{Z}^s$ of $\mathbb{R}^s/\mathbb{Z}^s$. If $\mathbf{h} \in \mathbb{Z}^s$, then*

$$\sum_{n=1}^{N} e^{2\pi \mathrm{i} \mathbf{h}\cdot\mathbf{y}_n} = \begin{cases} N & \text{for } \mathbf{h} \in L^{\perp}, \\ 0 & \text{for } \mathbf{h} \notin L^{\perp}. \end{cases}$$

Proof. The group $A := L/\mathbb{Z}^s$ is finite and also abelian as a subgroup of $\mathbb{R}^s/\mathbb{Z}^s$. Therefore we can talk about characters of A. For fixed $\mathbf{h} \in \mathbb{Z}^s$ we put

$$\chi_{\mathbf{h}}(\mathbf{y} + \mathbb{Z}^s) = e^{2\pi \mathrm{i} \mathbf{h}\cdot\mathbf{y}} \qquad \text{for all } \mathbf{y} \in L.$$

The map $\chi_{\mathbf{h}}$ is well defined since the right-hand side does not depend on the representative that we pick from the coset $\mathbf{y} + \mathbb{Z}^s$. Furthermore, $\chi_{\mathbf{h}}$ is obviously a character of the additive group A. Now we can write

$$\sum_{n=1}^{N} e^{2\pi \mathrm{i} \mathbf{h}\cdot\mathbf{y}_n} = \sum_{a \in A} \chi_{\mathbf{h}}(a).$$

The last sum, being a character sum for the finite abelian group A of order N, is equal to N if the character $\chi_{\mathbf{h}}$ is trivial and equal to 0 if $\chi_{\mathbf{h}}$ is nontrivial (compare with Theorem 1.3.34). Moreover, $\chi_{\mathbf{h}}$ is trivial if and only if $\mathbf{h} \cdot \mathbf{y} \in \mathbb{Z}$ for all $\mathbf{y} \in L$, that is, if and only if $\mathbf{h} \in L^{\perp}$. □

The following definition generalizes Definition 4.3.2 and the subsequent result generalizes Theorem 4.3.3. We avoid a trivial case by assuming from now on that a lattice point set contains at least two points.

Definition 4.3.20 For an s-dimensional integration lattice L with $N := |L/\mathbb{Z}^s| \geq 2$, we put

$$R(L) = \sum_{\mathbf{h} \in F(L)} \frac{1}{r(\mathbf{h})}$$

with $F(L) = C_s^*(N) \cap L^{\perp}$. We again use the convention that an empty sum is equal to 0.

Theorem 4.3.21 *Let L be an s-dimensional integration lattice and let $\mathcal{P}$ be the corresponding lattice point set with $N \geq 2$ points. Then*

$$D_N(\mathcal{P}) \leq \frac{s}{N} + \frac{1}{2} R(L).$$

Proof. Let $\mathbf{y}_1, \dots, \mathbf{y}_N$ be the points of $\mathcal{P}$. From the fact that the group $L/\mathbb{Z}^s$ has order N, it follows that $N\mathbf{y}_n \in \mathbb{Z}^s$ for $1 \leq n \leq N$. Therefore we can apply Proposition 4.3.1 with $M = N$, which yields

$$D_N(\mathcal{P}) \leq \frac{s}{N} + \sum_{\mathbf{h} \in C_s^*(N)} \frac{1}{r(\mathbf{h}, N)} \Big| \frac{1}{N} \sum_{n=1}^{N} e^{2\pi \mathrm{i} \mathbf{h}\cdot\mathbf{y}_n} \Big|.$$

Lemma 4.3.19 shows that

$$D_N(\mathcal{P}) \leq \frac{s}{N} + \sum_{\mathbf{h} \in F(L)} \frac{1}{r(\mathbf{h}, N)}.$$

The final step of the proof is the same as in the proof of Theorem 4.3.3. □

It should be quite obvious that there is a complete analog of Theorem 4.3.5 for general lattice rules, namely that for integrands $f \in C^k(\mathbb{R}^s/\mathbb{Z}^s)$ for some integer $k \geq 2$ and for an s-dimensional integration lattice L, the error in (4.44) satisfies the bound

$$\Big| \int_{[0,1]^s} f(\mathbf{u}) d\mathbf{u} - \frac{1}{N} \sum_{n=1}^{N} f(\mathbf{y}_n) \Big| \leq c(f) P_k(L),$$

where

$$P_k(L) = \sum_{\mathbf{h} \in L^\perp \setminus \{\mathbf{0}\}} r(\mathbf{h})^{-k}.$$

The proof of this bound is again based on Lemma 4.3.19. An analog of Proposition 4.3.6, that is, a bound on $P_k(L)$ in terms of $R(L)$, can be found in [134, Theorem 5.26].

For the proof of the following theorem, we need a notion from group theory. Let A be a finite abelian group with the additive notation and let G and H be subgroups of A. Then A is said to be the *direct sum* of G and H, written $A = G \oplus H$, if every element $a \in A$ can be written as $a = g + h$ with uniquely determined $g \in G$ and $h \in H$ (or equivalently, if for every $a \in A$ there is a representation $a = g + h$ with some $g \in G$ and $h \in H$ and if $G \cap H = \{0\}$). Each of G and H is called a *direct summand* of A. We proceed similarly for more than two direct summands, using the characterization in terms of the unique sum representation. We refer to Definition 1.3.22 for the concept of the exponent of A.

Lemma 4.3.22 *Let A be a finite abelian group and let $E = E(A)$ be the exponent of A. Then every cyclic subgroup of A of order E is a direct summand of A.*

Proof. We proceed by induction on the order of A. The case $A = \{0\}$ is trivial. Now we take a finite abelian group A of order greater than 1 and we assume that the lemma is already shown for all finite abelian groups of smaller order than A.

Let $C = \langle c \rangle$ be a cyclic subgroup of A of order E. If $C = A$, then $A = C \oplus \{0\}$ and C is a direct summand of A. If $C \neq A$, then we choose $b \in A \backslash C$ such that $\mathrm{ord}(b)$ is minimal among all elements of $A \setminus C$. Now $b \neq 0 \in A$ implies that $\mathrm{ord}(b) \geq 2$, and so we can talk about prime factors of $\mathrm{ord}(b)$. Take any prime factor p of $\mathrm{ord}(b)$ and consider the p-fold sum $d = pb$. Then $\mathrm{ord}(d) = \mathrm{ord}(b)/p < \mathrm{ord}(b)$, and in view of the definition of b we must have $d \in C$. Therefore $d = nc$ for some $n \in \mathbb{N}$. Now

$$\frac{\mathrm{ord}(b) n}{p} c = \frac{\mathrm{ord}(b)}{p}(nc) = \frac{\mathrm{ord}(b)}{p} d = 0 \in A,$$

and so E divides $\operatorname{ord}(b)n/p$ by Lemma 1.3.10. But $\operatorname{ord}(b)$ divides E by Proposition 1.3.24, and therefore p divides n, say $n = jp$ for some $j \in \mathbb{N}$. For the element $b - jc \notin C$ (recall that $b \notin C = \langle c \rangle$) we obtain

$$p(b - jc) = pb - (jp)c = d - nc = d - d = 0 \in A,$$

and so $\operatorname{ord}(b - jc) = p$. The minimality property of $\operatorname{ord}(b)$ implies that $\operatorname{ord}(b) \leq \operatorname{ord}(b - jc) = p$. Since p is a prime factor of $\operatorname{ord}(b)$, it follows that $\operatorname{ord}(b) = p$.

We introduce the cyclic subgroup $B = \langle b \rangle$ of A of order p. The intersection $B \cap C$ is a subgroup of B, and we deduce from Lagrange's theorem (see Theorem 1.3.21) and $b \notin C$ that $B \cap C = \{0\}$. Now we consider the factor group $\overline{A} := A/B$. Let m be the order of the element $c + B \in \overline{A}$. Then $m(c+B) = 0 \in \overline{A}$, that is, $mc + B = 0 + B$, and so $mc \in B$. But also $mc \in C$, hence $mc = 0 \in A$ since $B \cap C = \{0\}$, and so E divides m by Lemma 1.3.10. Furthermore $E(c + B) = Ec + B = 0 + B$, and so $m = E$.

Thus, we arrive at the following situation: $\overline{A}$ is a finite abelian group of smaller order than A, it contains the cyclic subgroup $\overline{C} := \langle c + B \rangle$ of order E, and E is the exponent of $\overline{A}$ since the exponent of a factor group of A cannot be larger than the exponent of A. Therefore we can apply the induction hypothesis to $\overline{A}$. This yields a subgroup $\overline{H}$ of $\overline{A}$ with $\overline{A} = \overline{C} \oplus \overline{H}$. Now $\overline{H}$ gives rise to the subgroup H of A that consists of all $h \in A$ with $h + B \in \overline{H}$. Note that $B \subseteq H$. It is then clear that every $a \in A$ can be written in the form $a = g + h$ with some $g \in C$ and $h \in H$. We obtain $A = C \oplus H$ if we can show that $C \cap H = \{0\}$. So let $tc \in H$ for some $t \in \mathbb{N}$. Then $tc + B \in \overline{H}$, but also $tc + B = t(c + B) \in \overline{C}$, and so $tc + B \in \overline{C} \cap \overline{H}$. This intersection consists only of the coset $0 + B$, and therefore $tc \in B$. This implies $tc \in B \cap C = \{0\}$ as desired. □

Now we return to the finite abelian group $L/\mathbb{Z}^s$ and we apply the theory of finite abelian groups, and in particular Lemma 4.3.22, in order to derive a canonical form of lattice point sets.

Theorem 4.3.23 *For every dimension $s \geq 1$ and every integer $N \geq 2$, an s-dimensional lattice point set with N points consists exactly of all fractional parts*

$$\Big\{ \sum_{i=1}^{r} (k_i/n_i)\mathbf{g}_i \Big\} \qquad \text{with } k_i \in \mathbb{Z},\ 0 \leq k_i < n_i \text{ for } 1 \leq i \leq r,$$

where the integer r with $1 \leq r \leq s$ and the integers $n_1, \ldots, n_r \geq 2$ with n_{i+1} dividing n_i for $1 \leq i \leq r - 1$ and $n_1 \cdots n_r = N$ are uniquely determined. Furthermore, the lattice points $\mathbf{g}_1, \ldots, \mathbf{g}_r \in \mathbb{Z}^s$ are linearly independent and for each $i = 1, \ldots, r$ the greatest common divisor of all coordinates of $\mathbf{g}_i$ and of n_i is equal to 1.

Proof. Let $A := L/\mathbb{Z}^s$ be the finite abelian group of order N corresponding to the given lattice point set. First we establish a suitable direct sum decomposition of A

by the following procedure. Let $n_1 \geq 2$ be the exponent of A and let C_1 be a cyclic subgroup of A of order n_1. If $C_1 = A$, then we stop. Otherwise, we apply Lemma 4.3.22 and obtain $A = C_1 \oplus A_2$ with a subgroup A_2 of A of order at least 2. Let $n_2 \geq 2$ be the exponent of A_2. Then n_2 divides n_1 by Proposition 1.3.24. Furthermore, A_2 has a cyclic direct summand C_2 of order n_2 by Lemma 4.3.22. Continuing in this way, we arrive after finitely many steps at a decomposition $A = C_1 \oplus \cdots \oplus C_r$, where C_i is a cyclic group of order $n_i \geq 2$ for $1 \leq i \leq r$ and n_{i+1} divides n_i for $1 \leq i \leq r-1$. A comparison of orders yields $N = n_1 \cdots n_r$. The number r and the orders $n_1, \ldots, n_r$ of the direct summands in this decomposition are uniquely determined by the multiset (that is, the set with multiplicities of elements taken into account) of orders of all elements of A.

For $i = 1, \ldots, r$, let $\mathbf{c}_i \in \mathbb{R}^s$ be such that $\mathbf{c}_i + \mathbb{Z}^s$ is a generator of the cyclic group C_i. Since C_i has order n_i, we have $n_i \mathbf{c}_i \in \mathbb{Z}^s$, and so $\mathbf{c}_i = (1/n_i)\mathbf{g}_i$ for some $\mathbf{g}_i \in \mathbb{Z}^s$. Since $\mathbf{c}_i + \mathbb{Z}^s \in C_i$ has order n_i, the greatest common divisor of all coordinates of $\mathbf{g}_i$ and of n_i is equal to 1. Furthermore, it follows from $A = C_1 \oplus \cdots \oplus C_r$ that the points of the given lattice point set are as indicated in the theorem.

If $\mathbf{c}_1, \ldots, \mathbf{c}_r$ were linearly dependent, then $\mathbf{0} \in \mathbb{R}^s$ could be written as a nontrivial linear combination of $\mathbf{c}_1, \ldots, \mathbf{c}_r$ with rational coefficients. By clearing denominators, we get $\sum_{i=1}^r j_i \mathbf{c}_i = \mathbf{0}$ with integers $j_1, \ldots, j_r$ not all 0 and satisfying $\gcd(j_1, \ldots, j_r) = 1$. This yields the identity

$$\sum_{i=1}^{r} j_i(\mathbf{c}_i + \mathbb{Z}^s) = \mathbf{0} + \mathbb{Z}^s$$

in the group A. The direct sum decomposition $A = C_1 \oplus \cdots \oplus C_r$ implies that $j_i(\mathbf{c}_i + \mathbb{Z}^s)$ is the identity element of C_i for $1 \leq i \leq r$, and so n_i divides j_i for $1 \leq i \leq r$. Since n_r divides all n_i, we infer that $n_r \geq 2$ divides $j_1, \ldots, j_r$. But this is a contradiction to $\gcd(j_1, \ldots, j_r) = 1$. Therefore $\mathbf{c}_1, \ldots, \mathbf{c}_r$, and so $\mathbf{g}_1, \ldots, \mathbf{g}_r$, are linearly independent. In particular, it follows that $r \leq s$. Now all claims in the theorem are proved. □

Definition 4.3.24 The uniquely determined integer r in Theorem 4.3.23 is called the *rank* of the lattice point set and the uniquely determined integers $n_1, \ldots, n_r$ in Theorem 4.3.23 are called the *invariants* of the lattice point set.

Example 4.3.25 Consider a point set $\mathcal{P}(\mathbf{g}, N)$ from Subsection 4.3.1 with $\mathbf{g} = (g_1, \ldots, g_s) \in \mathbb{Z}^s$ satisfying $\gcd(g_1, \ldots, g_s, N) = 1$. Then according to Definition 4.3.24, $\mathcal{P}(\mathbf{g}, N)$ is a lattice point set of rank 1 and its only invariant is $n_1 = N$. If $d := \gcd(g_1, \ldots, g_s, N) > 1$, then all points of $\mathcal{P}(\mathbf{g}, N)$ occur with the same multiplicity d, and so this case is not interesting in practice.

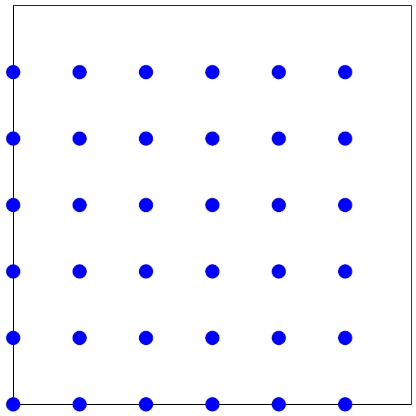

Figure 4.7: A shifted centered regular lattice

Example 4.3.26 Take the centered regular lattice in (4.33) with $m \geq 2$ and shift it so that the origin belongs to the shifted point set. The points of the shifted point set are

$$\left(\frac{k_1}{m}, \ldots, \frac{k_s}{m}\right) \in [0,1)^s$$

with $k_1, \ldots, k_s$ running independently through the integers $0, 1, \ldots, m-1$. We refer to Figure 4.7 for an illustration with $s = 2$ and $m = 6$. The corresponding subgroup A of $\mathbb{R}^s/\mathbb{Z}^s$ is obviously the direct sum $A = C \oplus \cdots \oplus C$ of s copies of a cyclic group C of order m. The uniqueness of the rank and of the invariants of lattice point sets implies that our lattice point set has rank s and invariants $n_1, \ldots, n_s$ with $n_i = m$ for $1 \leq i \leq s$. The lattice corresponding to this example is $L = (1/m)\mathbb{Z}^s$. The dual lattice of L is $L^\perp = m\mathbb{Z}^s$.

Example 4.3.27 The preceding example suggests the following general procedure. Let L be an s-dimensional integration lattice and let $m \geq 2$ be an integer. Then the scaled version $(1/m)L$ of L is again an s-dimensional integration lattice which can be thought of as a copy of L with scaling factor $1/m$. The corresponding lattice rule is called a *copy rule*. If $\mathbf{y}_1, \ldots, \mathbf{y}_N$ are the points of L in $[0,1)^s$, then all points of L are given by $\mathbf{y}_n + \mathbf{k}$ with $1 \leq n \leq N$ and $\mathbf{k}$ running through $\mathbb{Z}^s$. The points of $(1/m)L$ are therefore given by $(1/m)\mathbf{y}_n + (1/m)\mathbf{k}$ with $1 \leq n \leq N$ and $\mathbf{k}$ running through $\mathbb{Z}^s$. It is now easy to find the points of $(1/m)L$ belonging to $[0,1)^s$, and indeed the lattice point set corresponding to $(1/m)L$ consists of the m^sN points

$$\frac{1}{m}\mathbf{y}_n + \frac{1}{m}(k_1, \ldots, k_s) \in [0,1)^s,$$

where $1 \leq n \leq N$ and $k_1, \ldots, k_s$ run independently through the integers $0, 1, \ldots, m-1$. This point set can be viewed geometrically as follows: subdivide the s-dimensional unit cube into m^s smaller cubes each of side length $1/m$, and in each smaller cube choose an appropriately shifted and scaled-down version (by the factor $1/m$) of the points $\mathbf{y}_1, \ldots, \mathbf{y}_N$. Sloan and Joe [189, Sections 6.4 and 6.5] make a good case for the choice $m = 2$ which can improve the performance of certain lattice rules.

Much more is known about general lattice rules, and we refer you to [134, Chapter 5] and [189] if you want to learn more about them.

4.4 Nets and (t, s)-sequences

4.4.1 Basic facts about nets

We continue with the investigation of one of the key problems in quasi-Monte Carlo integration, namely the construction of low-discrepancy point sets. A rich source of low-discrepancy point sets is supplied by the theory of so-called nets or more precisely (t, m, s)-nets. The philosophy behind the concept of a net is very simple. In view of the definition of the discrepancy of a point set $\mathcal{P}$ consisting of N points in $[0, 1)^s$ (see Definition 4.1.35), it is clear that in order to arrive at a low-discrepancy point set $\mathcal{P}$, we need to make the counting function $A(J; \mathcal{P})$ roughly equal to $N\lambda_s(J)$ for all half-open subintervals J of $[0, 1)^s$. In an ideal world, $A(J; \mathcal{P})$ would be exactly equal to $N\lambda_s(J)$ for all such intervals J, but this is impossible because of the lower bounds on $D_N(\mathcal{P})$ stated in Subsection 4.1.2, or because of the even simpler reason that $A(J; \mathcal{P})$ is an integer and $N\lambda_s(J)$ is not always an integer. However, what is in fact feasible is to request the identity $A(J; \mathcal{P}) = N\lambda_s(J)$ for a large *finite* family of intervals J. The intuitive idea is then that if $A(J; \mathcal{P}) = N\lambda_s(J)$ for many intervals J, then $\mathcal{P}$ should overall be a low-discrepancy point set. This expectation is borne out by the results to be described below.

Some care has to be taken concerning the actual form of the intervals J for which we request that $A(J; \mathcal{P}) = N\lambda_s(J)$. The following examples provide a clue and lead to the notion introduced in Definition 4.4.3 below.

Example 4.4.1 For integers $b \geq 2$ and $m \geq 0$, consider the equidistant point set $\mathcal{P}$ consisting of the b^m rational numbers $0, 1/b^m, 2/b^m, \ldots, (b^m - 1)/b^m$ in $[0, 1)$. If we want $A(J; \mathcal{P}) = N\lambda_s(J) = b^m\lambda(J)$, then the length $\lambda(J)$ of the interval $J \subseteq [0, 1)$ must of course be a rational number with denominator b^m. The smallest intervals of this type have length b^{-m}. Indeed, every half-open interval $J_a := [ab^{-m}, (a+1)b^{-m})$ with $a \in \mathbb{Z}$ and $0 \leq a < b^m$ satisfies $A(J_a; \mathcal{P}) = b^m\lambda(J_a)$ since $A(J_a; \mathcal{P}) = 1$, that is, J_a contains exactly one point of the point set $\mathcal{P}$. Furthermore, the intervals J_a form a partition of $[0, 1)$ and any union J of intervals J_a satisfies again $A(J; \mathcal{P}) = b^m\lambda(J)$. The same holds if we replace $\mathcal{P}$ by the first b^m terms of the van der Corput sequence $\mathcal{S} = (\phi_b(n))_{n=0}^{\infty}$ in base b (see Remark 4.2.9), since the first b^m terms of $\mathcal{S}$ are just a rearrangement of the numbers $0, 1/b^m, 2/b^m, \ldots, (b^m - 1)/b^m$.

Example 4.4.2 For integers $b \geq 2$ and $m \geq 1$, let $\mathcal{P}$ be the two-dimensional Hammersley point set in (4.32) with $N = b^m$ and $b_1 = b$, that is, $\mathcal{P}$ consists of the b^m points

$$\mathbf{y}_n = \left(\frac{n}{b^m}, \phi_b(n)\right) \in [0,1)^2 \qquad \text{for } n = 0, 1, \ldots, b^m - 1.$$

Clearly, if $A(J;\mathcal{P}) = N\lambda_s(J) = b^m \lambda_2(J)$, then the area $\lambda_2(J)$ of the interval (in this case the rectangle) J must be a rational number with denominator b^m. The smallest rectangles of this type have area b^{-m}. In view of the b-adic nature of the points $\mathbf{y}_n$ of $\mathcal{P}$, it is natural to consider b-adic rectangles

$$J = [a_1 b^{-d_1}, (a_1+1)b^{-d_1}) \times [a_2 b^{-d_2}, (a_2+1)b^{-d_2}) \subseteq [0,1)^2 \tag{4.46}$$

with $a_1, a_2, d_1, d_2 \in \mathbb{Z}$, $d_1 \geq 0$, $d_2 \geq 0$, $0 \leq a_1 < b^{d_1}$, and $0 \leq a_2 < b^{d_2}$. The condition $\lambda_2(J) = b^{-m}$ means that $d_1 + d_2 = m$. We claim that each rectangle J in (4.46) with $d_1 + d_2 = m$ contains exactly one point of $\mathcal{P}$. For $n = 0, 1, \ldots, b^m - 1$, it is obvious that $\mathbf{y}_n \in J$ if and only if $a_1 b^{m-d_1} = a_1 b^{d_2} \leq n < (a_1+1)b^{d_2}$ and $\phi_b(n) \in [a_2 b^{-d_2}, (a_2+1)b^{-d_2})$. The last condition amounts to saying that the first d_2 b-adic digits of $\phi_b(n)$ are prescribed, or equivalently that in the digit expansion $n = \sum_{j=0}^{\infty} z_j(n) b^j$ of n in (4.30) the digits $z_0(n), z_1(n), \ldots, z_{d_2-1}(n)$ are prescribed. But in the range $a_1 b^{d_2} \leq n < (a_1+1)b^{d_2}$ there is exactly one value of n with these prescribed digits, and so we get indeed $A(J;\mathcal{P}) = 1 = b^m \lambda_2(J)$. The intervals J in (4.46) with fixed d_1 and d_2 form a partition of $[0,1)^2$ and any disjoint union J_1 of these intervals with $d_1 + d_2 = m$ satisfies again $A(J_1;\mathcal{P}) = b^m \lambda_2(J_1)$.

Definition 4.4.3 Let $b \geq 2$ and $s \geq 1$ be integers. A half-open subinterval J of $[0,1)^s$ of the form

$$J = \prod_{i=1}^{s} [a_i b^{-d_i}, (a_i+1)b^{-d_i}) \tag{4.47}$$

with $a_i, d_i \in \mathbb{Z}$, $d_i \geq 0$, and $0 \leq a_i < b^{d_i}$ for $1 \leq i \leq s$ is called an *elementary interval in base b*.

Now we can let the cat out of the bag: the idea behind the concept of a net $\mathcal{P}$ in base b is that we request that each elementary interval in base b with a prescribed volume gets the same share of points of $\mathcal{P}$. The following general definition of nets was introduced by Niederreiter [130], while special cases were considered earlier by Sobol' [190] and Faure [49].

Definition 4.4.4 Let $b \geq 2$ and $s \geq 1$ be integers and let t and m be integers with $0 \leq t \leq m$. A (t, m, s)*-net in base* b is a point set $\mathcal{P}$ consisting of b^m points in $[0,1)^s$ such that $A(J;\mathcal{P}) = b^m \lambda_s(J) = b^t$ for every elementary interval $J \subseteq [0,1)^s$ in base b with $\lambda_s(J) = b^{t-m}$.

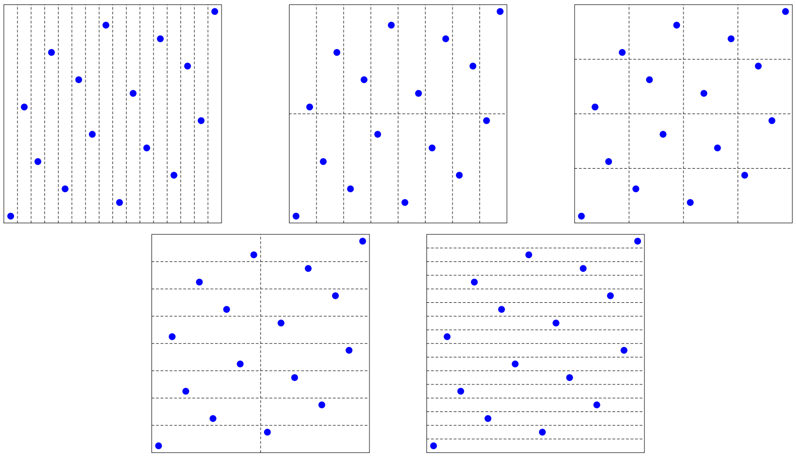

Figure 4.8: A $(0, 4, 2)$-net in base 2; every two-dimensional elementary interval in base 2 of area 2^{-4} contains exactly one point of the net

Example 4.4.5 The point set in Example 4.4.1 is a $(0, m, 1)$-net in base b. The point set in Example 4.4.2 is a $(0, m, 2)$-net in base b.

Example 4.4.6 Every point set of b^m points in $[0, 1)^s$ is a (t, m, s)-net in base b with $t = m$. For $m \geq 1$ we claim that we get a (t, m, s)-net in base b with $t = m - 1$ by taking the points

$$\left(\frac{n}{b}, \frac{n}{b}, \ldots, \frac{n}{b}\right) \in [0, 1)^s \qquad \text{for } n = 0, 1, \ldots, b - 1$$

on the main diagonal of the s-dimensional unit cube, each with multiplicity $b^t = b^{m-1}$. To begin with, this yields a point set $\mathcal{P}$ of b^m points in $[0, 1)^s$. According to Definition 4.4.4, we have to consider elementary intervals $J \subseteq [0, 1)^s$ in base b with $\lambda_s(J) = b^{t-m} = b^{-1}$. Because of the symmetry in the points of $\mathcal{P}$, it suffices to look at intervals J of the form $J = [ab^{-1}, (a+1)b^{-1}) \times [0, 1)^{s-1}$ with $a \in \mathbb{Z}$ and $0 \leq a < b$. It is clear that $(n/b, n/b, \ldots, n/b) \in J$ if and only if $n = a$. Therefore $A(J; \mathcal{P}) = b^t$ and $\mathcal{P}$ is indeed an $(m - 1, m, s)$-net in base b.

Example 4.4.7 The centered regular lattice in (4.33) can, in certain cases, be considered in a nontrivial way as a net. For integers $b \geq 2$, $r \geq 1$, and $s \geq 1$, let $\mathcal{P}$ be the point set consisting of the points

$$\mathbf{x}_{k_1,\ldots,k_s} = \left(\frac{2k_1 - 1}{2b^r}, \ldots, \frac{2k_s - 1}{2b^r}\right) \in [0, 1)^s$$

with $k_1, \ldots, k_s$ running independently through the integers $1, \ldots, b^r$. With $m = rs$ the number of points in $\mathcal{P}$ is $(b^r)^s = b^m$. We claim that $\mathcal{P}$ is an $(m - r, m, s)$-net in base b. Let $J \subseteq [0, 1)^s$ be an elementary interval in base b with $\lambda_s(J) = b^{-r}$. Then

J has the form (4.47) with $d_1 + \cdots + d_s = r$, and we must have $0 \leq d_i \leq r$ for $1 \leq i \leq s$. Now $\mathbf{x}_{k_1,\ldots,k_s} \in J$ if and only if

$$2k_i - 1 \in [2a_i b^{r-d_i}, (2a_i + 2)b^{r-d_i}) \qquad \text{for } 1 \leq i \leq s.$$

For fixed $i = 1, \ldots, s$, this condition holds exactly for $k_i = a_i b^{r-d_i} + 1, \ldots, a_i b^{r-d_i} + b^{r-d_i}$, that is, for exactly b^{r-d_i} values of k_i. Therefore

$$A(J; \mathcal{P}) = \prod_{i=1}^{s} b^{r-d_i} = b^{rs-(d_1+\cdots+d_s)} = b^{m-r},$$

and the claim is established.

Proposition 4.4.8 *Let $b \geq 2$, $s \geq 1$, and $0 \leq t \leq m$ be integers. If $\mathcal{P}$ is a (t, m, s)-net in base b, then $\mathcal{P}$ is also a (v, m, s)-net in base b for every integer v with $t \leq v \leq m$.*

Proof. It suffices to show that if $t < m$, then $\mathcal{P}$ is also a $(t + 1, m, s)$-net in base b. Let $J \subseteq [0, 1)^s$ be an elementary interval in base b with $\lambda_s(J) = b^{t+1-m}$. Then J has the form (4.47) with $d_1 + \cdots + d_s = m - t - 1$. For $c = 0, 1, \ldots, b - 1$, we introduce an elementary interval J_c in base b given by

$$J_c = [a_1 b^{-d_1} + c b^{-d_1-1}, a_1 b^{-d_1} + (c + 1)b^{-d_1-1}) \times \prod_{i=2}^{s} [a_i b^{-d_i}, (a_i + 1)b^{-d_i}).$$

Then $\lambda_s(J_c) = b^{t-m}$, and so $A(J_c; \mathcal{P}) = b^t$ for $0 \leq c \leq b - 1$ by the definition of a (t, m, s)-net in base b. Since J is the disjoint union of $J_0, J_1, \ldots, J_{b-1}$, we obtain

$$A(J; \mathcal{P}) = \sum_{c=0}^{b-1} A(J_c; \mathcal{P}) = b^{t+1}$$

as desired. □

Three of the four parameters t, m, s, and b of a (t, m, s)-net in base b are easy to determine: b is the base, s is the dimension, and m can be read off from the number of points in the net (which is b^m). The parameter t is also crucial since it tells us how small we can make an elementary interval J in base b and still get the perfect equidistribution property $A(J; \mathcal{P}) = b^m \lambda_s(J)$ in Definition 4.4.4. The number t is called the *quality parameter* of a (t, m, s)-net in base b. Definition 4.4.4 and Proposition 4.4.8 indicate that t should be small in order to get strong equidistribution properties of the net.

Proposition 4.4.8 is a simple instance of what is called a *propagation rule* for nets, that is, a rule that starts from one net or several nets and produces a net with new parameters. Here are two more propagation rules for nets that are simple but useful.

Proposition 4.4.9 *Let $b \geq 2$, $s \geq 2$, and $0 \leq t \leq m$ be integers and let r be an integer with $1 \leq r < s$. If $\mathcal{P}$ is a (t, m, s)-net in base b and $\mathcal{P}^{(r)}$ is as in Remark 4.1.37 the projection of $\mathcal{P}$ onto the first r coordinates, then $\mathcal{P}^{(r)}$ is a (t, m, r)-net in base b.*

Proof. The argument is similar to that in Remark 4.1.37. Let $J^{(r)} \subseteq [0,1)^r$ be an elementary interval in base b with $\lambda_r(J^{(r)}) = b^{t-m}$ and put $J = J^{(r)} \times [0,1)^{s-r} \subseteq [0,1)^s$. Then J is an elementary interval in base b with $\lambda_s(J) = b^{t-m}$. Since a projected point is in $J^{(r)}$ if and only if the original point is in J, we obtain $A(J^{(r)}; \mathcal{P}^{(r)}) = A(J; \mathcal{P}) = b^t$ by the definition of a (t, m, s)-net in base b, and so we are done. □

Proposition 4.4.10 *Let $b \geq 2$, $s \geq 1$, and $0 \leq t \leq m$ be integers. Then given a (t, m, s)-net in base b, we can construct a (t, k, s)-net in base b for every integer k with $t \leq k \leq m$.*

Proof. Let $\mathcal{P}$ be the given (t, m, s)-net in base b and fix an integer k with $t \leq k \leq m$. Consider the elementary interval $J_0 = [0, b^{k-m}) \times [0,1)^{s-1} \subseteq [0,1)^s$ in base b with $\lambda_s(J_0) = b^{k-m}$. Note that $\mathcal{P}$ is a (k, m, s)-net in base b by Proposition 4.4.8, and so $A(J_0; \mathcal{P}) = b^k$ by the definition of a (k, m, s)-net in base b. Let $\mathbf{x}_1, \ldots, \mathbf{x}_{b^k}$ be the points of $\mathcal{P}$ that belong to J_0. Let $\tau : J_0 \to [0,1)^s$ be the map defined by

$$\tau(u_1, u_2, \ldots, u_s) = (b^{m-k} u_1, u_2, \ldots, u_s) \qquad \text{for } (u_1, u_2, \ldots, u_s) \in J_0.$$

Now we claim that the point set $\mathcal{R}$ consisting of the points $\tau(\mathbf{x}_1), \ldots, \tau(\mathbf{x}_{b^k})$ is a (t, k, s)-net in base b. Take an elementary interval $J \subseteq [0,1)^s$ in base b with $\lambda_s(J) = b^{t-k}$. Then for $1 \leq n \leq b^k$, it is clear that $\tau(\mathbf{x}_n) \in J$ if and only if $\mathbf{x}_n \in \tau^{-1}(J) \subseteq J_0$, and furthermore $\tau^{-1}(J)$ is an elementary interval in base b with $\lambda_s(\tau^{-1}(J)) = b^{k-m} \lambda_s(J) = b^{t-m}$. It follows that $A(J; \mathcal{R}) = A(\tau^{-1}(J); \mathcal{P}) = b^t$ by the definition of a (t, m, s)-net in base b, and the proof is complete. □

The quality parameter t of a (t, m, s)-net in base b should be as small as possible in order to optimize the equidistribution properties of the net. Since t is, by definition, a nonnegative integer, the most favorable value of t is $t = 0$. This raises the question of whether we can always achieve $t = 0$ for any choice of the remaining parameters m, s, and b of a net. Unfortunately, the answer is *no*, and this provides further support for the conjecture that we are not living in the best of all possible worlds. In fact, the following theorem imposes a serious restriction on the existence of $(0, m, s)$-nets in base b.

Theorem 4.4.11 *Let $b \geq 2$ and $m \geq 2$ be integers. Then a $(0, m, s)$-net in base b can exist only if $s \leq b + 1$.*

Proof. We proceed by contradiction and assume that there exists a $(0, m, s)$-net in base b for some integers $m \geq 2$ and $s \geq b + 2$. Then Proposition 4.4.9 implies that there exists a $(0, m, b + 2)$-net in base b and Proposition 4.4.10 shows that there exists a $(0, 2, b + 2)$-net $\mathcal{P}$ in base b. Let $\mathbf{x}_1, \ldots, \mathbf{x}_{b^2}$ be the points of $\mathcal{P}$ and put

$$\mathbf{x}_n = \left(x_n^{(1)}, \ldots, x_n^{(b+2)}\right) \in [0, 1)^{b+2} \qquad \text{for } n = 1, \ldots, b^2.$$

For $i = 1, \ldots, b + 2$ and $n = 1, \ldots, b^2$, we set

$$a_n^{(i)} = \lfloor b x_n^{(i)} \rfloor \in Z_b = \{0, 1, \ldots, b - 1\}.$$

We take these integers $a_n^{(i)}$ and form the $(b + 2) \times b^2$ array

$$\begin{array}{llll} a_1^{(1)} & a_2^{(1)} & \ldots & a_{b^2}^{(1)} \\ a_1^{(2)} & a_2^{(2)} & \ldots & a_{b^2}^{(2)} \\ \vdots & \vdots & & \vdots \\ a_1^{(b+2)} & a_2^{(b+2)} & \ldots & a_{b^2}^{(b+2)} \end{array}$$

Now we consider a pair of rows of this array, say the ith row and the jth row with $1 \leq i < j \leq b+2$. For a given ordered pair $(z_1, z_2) \in Z_b^2$, we have $(a_n^{(i)}, a_n^{(j)}) = (z_1, z_2)$ if and only if $x_n^{(i)} \in [z_1/b, (z_1 + 1)/b)$ and $x_n^{(j)} \in [z_2/b, (z_2 + 1)/b)$, that is, if and only if $\mathbf{x}_n$ lies in the interval $J = \prod_{k=1}^{b+2} I_k \subseteq [0, 1)^{b+2}$ with $I_i = [z_1/b, (z_1 + 1)/b)$, $I_j = [z_2/b, (z_2 + 1)/b)$, and $I_k = [0, 1)$ for $k \in \{1, \ldots, b + 2\} \setminus \{i, j\}$. Now J is an elementary interval in base b with $\lambda_{b+2}(J) = b^{-2}$, and so the definition of a $(0, 2, b + 2)$-net in base b implies that $A(J; \mathcal{P}) = b^2 \lambda_{b+2}(J) = 1$. In other words, there is a one-to-one correspondence between the ordered pairs $(a_n^{(i)}, a_n^{(j)})$ from Z_b^2 and the integers $n = 1, \ldots, b^2$. The b^2 ordered pairs $(a_n^{(i)}, a_n^{(j)})$, $n = 1, \ldots, b^2$, run exactly through Z_b^2. We express this by saying that the ith row and the jth row of our array are orthogonal. Since this holds for any distinct i and j, we use the terminology that the rows of our array are mutually orthogonal.

It is a consequence of the mutual orthogonality of the rows of our array that each element of Z_b occurs exactly b times in each row of the array. Now we normalize the array in the following way. For each $i = 1, \ldots, b + 2$, we choose a permutation ψ_i of Z_b such that $\psi_i(a_1^{(i)}) = 0$, and then we transform the ith row $(a_1^{(i)} \ \ldots \ a_{b^2}^{(i)})$ of the array into the row $(\psi_i(a_1^{(i)}) \ \ldots \ \psi_i(a_{b^2}^{(i)}))$, that is, we apply ψ_i to each entry of the ith row. This amounts to a renaming of the elements, and so the mutual orthogonality of the rows of the array is preserved. Furthermore, in the new normalized array all entries in the first column are equal to 0.

Finally, we take the normalized array, delete its first column, and thus obtain a $(b + 2) \times (b^2 - 1)$ subarray. In each row of this subarray, exactly $b - 1$ entries are equal to 0, and so the total number of 0's in the subarray is $(b + 2)(b - 1)$. On the other hand, consider any of the $b^2 - 1$ columns of the subarray and suppose that it contains the element 0, say in the ith row. If there were a second entry 0 in this column, say in the jth row with $j \neq i$, then this would violate the orthogonality of the ith row and the jth row of the normalized array (recall that the first entry in each row of the normalized array is equal to 0). Hence each of the $b^2 - 1$ columns of the $(b+2) \times (b^2-1)$ subarray contains at most one entry equal to 0. It follows that the total number of 0's in the subarray is $\leq b^2 - 1 = (b+1)(b-1) < (b+2)(b-1)$, which is the desired contradiction. □

Remark 4.4.12 The condition $m \geq 2$ is needed for the validity of Theorem 4.4.11 since Example 4.4.6 shows that there exists a $(0, 0, s)$-net in base b and also a $(0, 1, s)$-net in base b for every $b \geq 2$ and every dimension $s \geq 1$.

Remark 4.4.13 Theorem 4.4.11 can be refined for various values of b by using the combinatorial theory of latin squares. A *latin square* of order $b \geq 2$ is a $b \times b$ array of elements from Z_b (or from any other set with b elements) such that each row and each column is a permutation of Z_b. A well-known example from the puzzle pages of newspapers is a sudoku which is a latin square of order 9 with Z_9 replaced by $\{1, \dots, 9\}$ and with additional requirements (see Figure 4.9). Two latin squares $S_1 = (s_{ij}^{(1)})_{1 \leq i,j \leq b}$ and $S_2 = (s_{ij}^{(2)})_{1 \leq i,j \leq b}$ of order b are *orthogonal* if the b^2 ordered pairs $(s_{ij}^{(1)}, s_{ij}^{(2)}) \in Z_b^2$, $i, j = 1, \dots, b$, are all distinct. A collection $S_1, \dots, S_k$ of latin squares of order b is *mutually orthogonal* if S_g and S_h are orthogonal for all $1 \leq g < h \leq k$. There is a maximum cardinality for a collection of mutually orthogonal latin squares of order b and this maximum cardinality is denoted by $M(b)$. Then it was proved in [130] that if $b \geq 2$ and $m \geq 2$ are integers, then a $(0, m, s)$-net in base b can exist only if $s \leq M(b) + 2$. Since $M(b) \leq b - 1$ for all $b \geq 2$, Theorem 4.4.11 is a consequence of the result in [130]. We have $M(b) = b - 1$ if b is a prime power, but there are values of b for which $M(b)$ is unexpectedly small, for instance $M(b) = 1$ for $b = 6$. Thus, for $m \geq 2$ a $(0, m, s)$-net in base 6 can exist only if $s \leq 3$. A book on latin squares that includes the connection with nets was written by Laywine and Mullen [94].

The first discrepancy bound for general (t, m, s)-nets in base b was established in [130] (see also [134, Theorem 4.10]), and it shows that a (t, m, s)-net in base b is a low-discrepancy point set if t is small compared to m, for instance if $t = 0$. Various improvements on the constants in this discrepancy bound were achieved later. We state without proof the following discrepancy bound which is obtained by combining results from [89] and [50].

Theorem 4.4.14 *Let $b \geq 2$, $s \geq 1$, and $m \geq 1$ be integers and let t be an integer with $0 \leq t \leq m$. Then the star discrepancy $D_N^*(\mathcal{P})$ of a (t, m, s)-net $\mathcal{P}$ in base b with $N = b^m$ satisfies*

$$ND_N^*(\mathcal{P}) \leq \frac{\lfloor b^2/2 \rfloor}{b^2 - 1} \cdot \frac{b^t}{(s-1)!} \Big(\frac{b-1}{2 \log b}\Big)^{s-1} (\log N)^{s-1} + B(b, s) b^t (\log N)^{s-2},$$

9	2	4	3	8	6	1	5	7
3	8	5	7	1	2	4	6	9
6	7	1	4	5	9	2	3	8
2	1	9	8	4	3	5	7	6
7	5	3	6	9	1	8	2	4
8	4	6	5	2	7	3	9	1
4	9	8	2	6	5	7	1	3
5	6	7	1	3	8	9	4	2
1	3	2	9	7	4	6	8	5

Figure 4.9: A sudoku, a latin square of order 9

where the constant $B(b, s) > 0$ *depends only on* b *and* s.

It is again evident from Theorem 4.4.14 that we prefer small values of the quality parameter t in a (t, m, s)-net in base b. This is in conformity with an earlier observation that smaller values of t imply stronger equidistribution properties of a (t, m, s)-net in base b. Because of the exponential dependence on t of the discrepancy bound in Theorem 4.4.14, even a small decrease in the value of t yields a considerable payoff in the discrepancy bound. Therefore it is worthwhile to work hard on the minimization of the value of t.

4.4.2 Digital nets and duality theory

Apart from some simple illustrations of the concept of a net in the preceding subsection, we have not yet seen concrete examples of good nets in arbitrary dimensions. What is still lacking in our presentation is an effective general instrument for the construction of nets. Such a tool is available in the case where the base b is a prime power, and in agreement with earlier practice in this book we write then q for the prime power. It should not come as a surprise that the reason why prime-power bases are special is that for a prime power q there exists a finite field with q elements, or of order q in the terminology of Section 1.4. The construction principle for nets that we will describe in the following is called the *digital method* and it is based on the theory of vector spaces and matrices over finite fields. We refer to Subsections 3.2.1 and 3.2.3 for a brief account of this theory. For the sake of completeness, it should be mentioned that versions of the digital method are available also for bases that are not prime powers (see [130] and [134, Section 4.3]), but the method is much more powerful for prime-power bases.

We focus on a simplified version of the digital method for prime-power bases and refer to [130] and [134, Section 4.3] for the more general original version. Let q be an arbitrary prime power and let $\mathbb{F}_q$ be the finite field of order q. Let $s \geq 1$ be a given dimension and let $m \geq 1$ be an integer. In order to obtain a (t, m, s)-net in base q, we have to construct q^m suitable points in $[0,1)^s$. The crucial step in the construction is to choose $m \times m$ matrices $C^{(1)}, \ldots, C^{(s)}$ over $\mathbb{F}_q$, that is, one matrix for each of the s coordinate directions of points in $[0,1)^s$. Next we set up the map $T_m : \mathbb{F}_q^m \to [0,1)$ by putting

$$T_m(\mathbf{h}) = \sum_{j=1}^{m} \psi(h_j) q^{-j} \tag{4.48}$$

for every column vector $\mathbf{h} = (h_1, \ldots, h_m)^\top \in \mathbb{F}_q^m$, where $\psi : \mathbb{F}_q \to Z_q$ is a fixed bijection from $\mathbb{F}_q$ onto the least residue system Z_q modulo q. For each column vector $\mathbf{v} \in \mathbb{F}_q^m$, we compute the matrix-vector products $C^{(i)}\mathbf{v} \in \mathbb{F}_q^m$ for $1 \le i \le s$, and then we associate with the vector $\mathbf{v}$ the point

$$(T_m(C^{(1)}\mathbf{v}), \ldots, T_m(C^{(s)}\mathbf{v})) \in [0,1)^s. \tag{4.49}$$

By letting $\mathbf{v}$ range over all q^m possibilities in $\mathbb{F}_q^m$, we arrive at a point set consisting of q^m points in $[0,1)^s$.

Definition 4.4.15 The point set $\mathcal{P}$ consisting of the q^m points in (4.49) is called a *digital net over* $\mathbb{F}_q$. If $\mathcal{P}$ forms a (t,m,s)-net in base q for some integer t with $0 \le t \le m$, then $\mathcal{P}$ is called a *digital* (t,m,s)*-net over* $\mathbb{F}_q$. The matrices $C^{(1)}, \ldots, C^{(s)}$ are the *generating matrices* of $\mathcal{P}$.

Example 4.4.16 Let $s = 1$, let q be an arbitrary prime power, and let $m \ge 1$ be an integer. We choose $C^{(1)}$ to be the $m \times m$ identity matrix over $\mathbb{F}_q$. Then for any bijection $\psi : \mathbb{F}_q \to Z_q$ in (4.48), the corresponding digital net $\mathcal{P}$ over $\mathbb{F}_q$ agrees with the equidistant point set in Example 4.4.1 for $b = q$. This point set $\mathcal{P}$ is a $(0,m,1)$-net in base q by Example 4.4.5, and so $\mathcal{P}$ is a digital $(0,m,1)$-net over $\mathbb{F}_q$ and $C^{(1)}$ is its generating matrix.

Example 4.4.17 Let $s = 2$, let q be an arbitrary prime power, and let $m \ge 1$ be an integer. We choose $C^{(1)}$ to be the $m \times m$ identity matrix over $\mathbb{F}_q$. Let $C^{(2)} = (c_{ij})_{1 \le i,j \le m}$ be the $m \times m$ antidiagonal matrix over $\mathbb{F}_q$ with $c_{ij} = 1$ if $i+j = m+1$ and $c_{ij} = 0$ otherwise. Then for any bijection $\psi : \mathbb{F}_q \to Z_q$ in (4.48), the corresponding digital net $\mathcal{P}$ agrees with the point set in Example 4.4.2 for $b = q$. This point set $\mathcal{P}$ is a $(0,m,2)$-net in base q by Example 4.4.5, and so $\mathcal{P}$ is a digital $(0,m,2)$-net over $\mathbb{F}_q$ and $C^{(1)}$ and $C^{(2)}$ are its generating matrices.

We know from Example 4.4.6 that $t = m$ is always a possible value of the quality parameter for a digital net over $\mathbb{F}_q$ consisting of q^m points. Hence every s-dimensional digital net over $\mathbb{F}_q$ with q^m points is a digital (t,m,s)-net over $\mathbb{F}_q$ for some value of t. We want to figure out by all means how we can obtain values of t smaller than m. It transpires that the quality parameter of a digital net depends only on its generating matrices and in fact on a certain linear independence property of the rows of the generating matrices. The following definition is convenient in this context.

Definition 4.4.18 Let q be a prime power, let $m \geq 1$ and $s \geq 1$ be integers, and let d be an integer with $0 \leq d \leq m$. The system $\{\mathbf{h}_j^{(i)} \in \mathbb{F}_q^m : 1 \leq j \leq m,\ 1 \leq i \leq s\}$ of vectors is a (d,m,s)-*system over* $\mathbb{F}_q$ if for all nonnegative integers $d_1, \ldots, d_s$ with $\sum_{i=1}^s d_i = d$, the system $\{\mathbf{h}_j^{(i)} \in \mathbb{F}_q^m : 1 \leq j \leq d_i,\ 1 \leq i \leq s\}$ is linearly independent over $\mathbb{F}_q$. Here an empty system (which occurs for $d = 0$) is considered linearly independent over $\mathbb{F}_q$.

It is clear from this definition that the property of being a (d,m,s)-system over $\mathbb{F}_q$ is the stronger the larger the value of d. We consider now the $m \times m$ generating matrices $C^{(1)}, \ldots, C^{(s)}$ over $\mathbb{F}_q$ of a digital net over $\mathbb{F}_q$. For $1 \leq i \leq s$ and $1 \leq j \leq m$, let $\mathbf{c}_j^{(i)} \in \mathbb{F}_q^m$ denote the jth row vector of the matrix $C^{(i)}$.

Theorem 4.4.19 *Let q be a prime power, let $m \geq 1$ and $s \geq 1$ be integers, and let t be an integer with $0 \leq t \leq m$. Then a digital net over $\mathbb{F}_q$ with $m \times m$ generating matrices $C^{(1)}, \ldots, C^{(s)}$ over $\mathbb{F}_q$ is a digital (t,m,s)-net over $\mathbb{F}_q$ if and only if the system $\{\mathbf{c}_j^{(i)} \in \mathbb{F}_q^m : 1 \leq j \leq m,\ 1 \leq i \leq s\}$ of row vectors of $C^{(1)}, \ldots, C^{(s)}$ is a (d,m,s)-system over $\mathbb{F}_q$ with $d = m - t$.*

Proof. Let

$$J = \prod_{i=1}^{s} [a_i q^{-d_i}, (a_i + 1) q^{-d_i}) \subseteq [0,1)^s$$

be an elementary interval in base q of the form (4.47) with $b = q$. Put $d = \sum_{i=1}^s d_i$, so that $\lambda_s(J) = q^{-d}$. We can assume that $d \geq 1$, for otherwise we have the trivial case $J = [0,1)^s$. For a column vector $\mathbf{v} \in \mathbb{F}_q^m$, the corresponding point in (4.49) lies in J if and only if

$$T_m(C^{(i)}\mathbf{v}) \in [a_i q^{-d_i}, (a_i + 1) q^{-d_i}) \qquad \text{for } 1 \leq i \leq s.$$

This condition means that for $1 \leq i \leq s$ the first d_i q-adic digits of $T_m(C^{(i)}\mathbf{v})$ and $a_i q^{-d_i}$ agree, and this amounts to saying that $C_{d,m}\mathbf{v} = \mathbf{b}$ for some column vector $\mathbf{b} \in \mathbb{F}_q^d$ depending only on J, where $C_{d,m}$ is a $d \times m$ matrix over $\mathbb{F}_q$ whose row vectors are given by the row vectors $\mathbf{c}_j^{(i)}$, $1 \leq j \leq d_i$, $1 \leq i \leq s$, of the generating matrices in some order. If the given digital net is a digital (t,m,s)-net over $\mathbb{F}_q$, then with $d = m - t$ the equation $C_{d,m}\mathbf{v} = \mathbf{b}$ has exactly q^t solutions $\mathbf{v} \in \mathbb{F}_q^m$ for every $\mathbf{b} \in \mathbb{F}_q^d$. This implies that the system $\{\mathbf{c}_j^{(i)} \in \mathbb{F}_q^m : 1 \leq j \leq d_i,\ 1 \leq i \leq s\}$ is linearly independent over $\mathbb{F}_q$, and since this holds for all choices of nonnegative integers $d_1, \ldots, d_s$ with $\sum_{i=1}^s d_i = m - t = d$, we infer that the system $\{\mathbf{c}_j^{(i)} \in \mathbb{F}_q^m : 1 \leq j \leq m,\ 1 \leq i \leq s\}$ is a (d,m,s)-system over $\mathbb{F}_q$ with $d = m - t$.

The converse holds for the same reasons. □

If we want to minimize the value of t for given generating matrices of a digital net, then according to Theorem 4.4.19 we have to maximize the value of d, and this leads naturally to the following concept.

Definition 4.4.20 The *figure of merit* $\varrho(C^{(1)},\dots,C^{(s)})$ of $m\times m$ matrices $C^{(1)},\dots,C^{(s)}$ over $\mathbb{F}_q$ is defined to be the largest integer d such that the system of row vectors of $C^{(1)},\dots,C^{(s)}$ is a (d,m,s)-system over $\mathbb{F}_q$.

Corollary 4.4.21 *A digital net over $\mathbb{F}_q$ with $m\times m$ generating matrices $C^{(1)},\dots,C^{(s)}$ over $\mathbb{F}_q$ is a digital (t,m,s)-net over $\mathbb{F}_q$ with*

$$t = m - \varrho(C^{(1)},\dots,C^{(s)}),$$

and this is the least value of t for this digital net.

Proof. This follows from Theorem 4.4.19 and Definition 4.4.20. □

It is an obvious consequence of Definitions 4.4.18 and 4.4.20 that the figure of merit $\varrho(C^{(1)},\dots,C^{(s)})$ of $m\times m$ matrices $C^{(1)},\dots,C^{(s)}$ over $\mathbb{F}_q$ always satisfies $0\le \varrho(C^{(1)},\dots,C^{(s)})\le m$. Corollary 4.4.21 signalizes that we should design generating matrices with a large figure of merit in order to obtain digital nets with a small quality parameter.

Example 4.4.22 Let $\mathcal{P}$ be the digital net over $\mathbb{F}_q$ in Example 4.4.16. Its generating matrix $C^{(1)}$ is the $m\times m$ identity matrix over $\mathbb{F}_q$. Since the row vectors of $C^{(1)}$ are linearly independent over $\mathbb{F}_q$, it is evident that $\varrho(C^{(1)}) = m$. Therefore Corollary 4.4.21 implies that $\mathcal{P}$ is a digital $(0,m,1)$-net over $\mathbb{F}_q$, and this agrees with the result in Example 4.4.16.

Example 4.4.23 Consider the digital net $\mathcal{P}$ over $\mathbb{F}_q$ in Example 4.4.17 with the generating matrices $C^{(1)}$ and $C^{(2)}$ stipulated there. We claim that $\varrho(C^{(1)},C^{(2)}) = m$. For $k=1,\dots,m$, let $\mathbf{s}_k$ be the kth unit vector in $\mathbb{F}_q^m$, that is, the vector with kth coordinate equal to 1 and all other coordinates equal to 0. Now we take two integers $d_1\ge 0$ and $d_2\ge 0$ with $d_1+d_2=m$. The first d_1 row vectors of $C^{(1)}$ are the unit vectors $\mathbf{s}_1,\mathbf{s}_2,\dots,\mathbf{s}_{d_1}$ and the first $d_2=m-d_1$ row vectors of $C^{(2)}$ are the unit vectors $\mathbf{s}_m,\mathbf{s}_{m-1},\dots,\mathbf{s}_{d_1+1}$. Therefore the system $\{\mathbf{c}_j^{(i)}\in\mathbb{F}_q^m : 1\le j\le d_i,\ 1\le i\le 2\}$ of row vectors of $C^{(1)}$ and $C^{(2)}$ consists exactly of all unit vectors $\mathbf{s}_1,\dots,\mathbf{s}_m$ in $\mathbb{F}_q^m$, and these unit vectors are obviously linearly independent over $\mathbb{F}_q$. Thus, we have indeed $\varrho(C^{(1)},C^{(2)}) = m$. Hence Corollary 4.4.21 implies that $\mathcal{P}$ is a digital $(0,m,2)$-net over $\mathbb{F}_q$, in accordance with the result in Example 4.4.17.

Remark 4.4.24 It follows from Theorem 4.4.11 and Corollary 4.4.21 that for integers $m \geq 2$ and $s \geq q+2$, there cannot exist $m \times m$ matrices $C^{(1)}, \ldots, C^{(s)}$ over $\mathbb{F}_q$ with figure of merit $\varrho(C^{(1)}, \ldots, C^{(s)}) = m$. This can be proved also directly by the theory of vector spaces. If there were such matrices $C^{(1)}, \ldots, C^{(s)}$, then their row vectors would form an (m, m, s)-system $S = \{\mathbf{c}_j^{(i)} \in \mathbb{F}_q^m : 1 \leq j \leq m,\ 1 \leq i \leq s\}$ over $\mathbb{F}_q$. Then $\{\mathbf{c}_1^{(1)}, \ldots, \mathbf{c}_m^{(1)}\}$ is a basis of the vector space $\mathbb{F}_q^m$. In the representation of each vector $\mathbf{c}_1^{(i)}$, $2 \leq i \leq s$, as a linear combination of these basis vectors, the coefficient of $\mathbf{c}_m^{(1)}$ must be nonzero by the definition of an (m, m, s)-system over $\mathbb{F}_q$. Thus for each $i = 2, \ldots, s$, there exists a nonzero $f_i \in \mathbb{F}_q$ such that $f_i \mathbf{c}_1^{(i)} - \mathbf{c}_m^{(1)}$ is a linear combination of $\mathbf{c}_1^{(1)}, \ldots, \mathbf{c}_{m-1}^{(1)}$. Let $b_i \in \mathbb{F}_q$ be the coefficient of $\mathbf{c}_{m-1}^{(1)}$ in the last linear combination. Since $s \geq q+2$, two of the elements $b_2, \ldots, b_s$ of $\mathbb{F}_q$ must be identical, say $b_h = b_k$ for some subscripts h and k with $2 \leq h < k \leq s$. Then by subtraction we see that $f_h \mathbf{c}_1^{(h)} - f_k \mathbf{c}_1^{(k)}$ is a linear combination of $\mathbf{c}_1^{(1)}, \ldots, \mathbf{c}_{m-2}^{(1)}$ (or equal to $\mathbf{0} \in \mathbb{F}_q^m$ if $m = 2$), and this is a contradiction to S being an (m, m, s)-system over $\mathbb{F}_q$.

Remark 4.4.25 At this stage we can already observe some connections between digital nets and linear codes. Let C be a linear $[s, k]$ code over $\mathbb{F}_q$ with $1 \leq k \leq s-1$ and with minimum distance $d(C) \geq d+1$ for some integer $d \geq 1$. Then a parity-check matrix H of C is an $(s-k) \times s$ matrix over $\mathbb{F}_q$, say with column vectors $\mathbf{c}^{(1)}, \ldots, \mathbf{c}^{(s)}$. Hence by Theorem 3.2.44, any d of the vectors $\mathbf{c}^{(1)}, \ldots, \mathbf{c}^{(s)}$ are linearly independent over $\mathbb{F}_q$. Thus, the construction of a good linear code obliges us to find a list of s vectors $\mathbf{c}^{(1)}, \ldots, \mathbf{c}^{(s)}$, that is, a $1 \times s$ array of vectors, with the indicated linear independence property for a large value of d. The construction of a good digital (t, m, s)-net over $\mathbb{F}_q$ challenges us to find an $m \times s$ array of vectors $\mathbf{c}_j^{(i)} \in \mathbb{F}_q^m$, $1 \leq j \leq m$, $1 \leq i \leq s$, with the linear independence property captured by the definition of a (d, m, s)-system over $\mathbb{F}_q$ with a large value of d (see Definition 4.4.18 and Theorem 4.4.19). In this sense, we can think of the construction of good digital nets over $\mathbb{F}_q$ as being a harder problem than the construction of good linear codes over $\mathbb{F}_q$. We will see further links between digital nets and linear codes in the duality theory for digital nets described below and later on in Theorem 4.4.35.

A basic fact about linear codes is the connection between minimum distance and Hamming weights in Theorem 3.2.14. There is an equally fundamental relationship between the quality parameter of a digital net and generalizations of Hamming weights, and this is the pivot of the duality theory for digital nets developed by Niederreiter and Pirsic [139].

Let q be a prime power and let $m \geq 1$ be an integer. We introduce a weight function v_m on $\mathbb{F}_q^m$ by putting $v_m(\mathbf{b}) = 0$ if $\mathbf{b} = \mathbf{0} \in \mathbb{F}_q^m$, and for $\mathbf{b} = (b_1, \ldots, b_m) \in \mathbb{F}_q^m$ with $\mathbf{b} \neq \mathbf{0}$ we let $v_m(\mathbf{b})$ be the largest value of j with $1 \leq j \leq m$ such that $b_j \neq 0$.

Definition 4.4.26 Let q be a prime power and let $m \geq 1$ and $s \geq 1$ be integers. Write a vector $\mathbf{B} \in \mathbb{F}_q^{ms}$ as the concatenation of s vectors of length m, that is, $\mathbf{B} = (\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(s)}) \in \mathbb{F}_q^{ms}$ with $\mathbf{b}^{(i)} \in \mathbb{F}_q^m$ for $1 \leq i \leq s$. Then the *NRT weight* $V_m(\mathbf{B})$ of $\mathbf{B}$ is defined by

$$V_m(\mathbf{B}) = \sum_{i=1}^{s} v_m(\mathbf{b}^{(i)}).$$

The NRT weight is named after the work of Niederreiter, Rosenbloom, and Tsfasman. The NRT weight was first introduced by Niederreiter [129] in the context of research on low-discrepancy point sets, and it was later applied in coding theory by Rosenbloom and Tsfasman [174]. If the distance $d_m(\mathbf{A}, \mathbf{B})$ of $\mathbf{A}, \mathbf{B} \in \mathbb{F}_q^{ms}$ is defined by $d_m(\mathbf{A}, \mathbf{B}) = V_m(\mathbf{A} - \mathbf{B})$, then the pair $(\mathbb{F}_q^{ms}, d_m)$ forms a metric space (compare with Remark 3.1.8) called the *NRT space*. For $m = 1$ the NRT space reduces to the Hamming space $(\mathbb{F}_q^s, d_1)$ with d_1 being the Hamming distance on $\mathbb{F}_q^s$ (see again Remark 3.1.8).

Example 4.4.27 Let $q = 2$, $m = 5$, $s = 2$, and consider the vector

$$\mathbf{B} = (0, 0, 1, 1, 0, 1, 0, 1, 0, 0) \in \mathbb{F}_2^{10}.$$

Then $\mathbf{B} = (\mathbf{b}^{(1)}, \mathbf{b}^{(2)})$ with $\mathbf{b}^{(1)} = (0, 0, 1, 1, 0) \in \mathbb{F}_2^5$ and $\mathbf{b}^{(2)} = (1, 0, 1, 0, 0) \in \mathbb{F}_2^5$. Clearly $v_5(\mathbf{b}^{(1)}) = 4$ and $v_5(\mathbf{b}^{(2)}) = 3$, and therefore $V_5(\mathbf{B}) = v_5(\mathbf{b}^{(1)}) + v_5(\mathbf{b}^{(2)}) = 7$ by Definition 4.4.26. Note that if we keep the same vector $\mathbf{B}$, but change m and s to $m = 2$ and $s = 5$, then an easy computation shows that $V_2(\mathbf{B}) = 6$. Thus, on the same vector space $\mathbb{F}_q^{ms}$ the NRT weight function V_m depends on m, and this is why we write m in the subscript of V.

The following definition of minimum distance is inspired by Theorem 3.2.14 in coding theory. For the reason given in Example 4.4.27, it is important to point out the dependence on m in the notation for this minimum distance.

Definition 4.4.28 Let q be a prime power and let $m \geq 1$ and $s \geq 1$ be integers. Then the *minimum distance* $\delta_m(\mathcal{N})$ of a nonzero linear subspace $\mathcal{N}$ of $\mathbb{F}_q^{ms}$ is defined by

$$\delta_m(\mathcal{N}) = \min_{\mathbf{B} \in \mathcal{N} \setminus \{\mathbf{0}\}} V_m(\mathbf{B}).$$

It is trivial that always $\delta_m(\mathcal{N}) \geq 1$. As to an upper bound, it is remarkable that the classical Singleton bound for linear codes (see Corollary 3.4.11) can be generalized to the minimum distance $\delta_m(\mathcal{N})$. The Singleton bound corresponds to the case $m = 1$ in the following proposition. As in Chapter 3, we write $\dim(\mathcal{N})$ for the dimension of a finite-dimensional vector space $\mathcal{N}$ over $\mathbb{F}_q$.

Proposition 4.4.29 *Let q be a prime power and let $m \geq 1$ and $s \geq 1$ be integers. Then every nonzero linear subspace $\mathcal{N}$ of $\mathbb{F}_q^{ms}$ satisfies*

$$\delta_m(\mathcal{N}) \leq ms - \dim(\mathcal{N}) + 1.$$

Proof. Put $k = \dim(\mathcal{N})$ and let $\pi : \mathcal{N} \to \mathbb{F}_q^k$ be the linear transformation that maps $\mathbf{B} \in \mathcal{N}$ to the k-tuple of the last k coordinates of $\mathbf{B}$. If π is surjective, then there exists a nonzero $\mathbf{B}_1 \in \mathcal{N}$ with

$$\pi(\mathbf{B}_1) = (1, 0, \ldots, 0) \in \mathbb{F}_q^k.$$

Then $V_m(\mathbf{B}_1) \leq ms - k + 1$. If π is not surjective, then it follows from $\dim(\mathcal{N}) = \dim(\mathbb{F}_q^k)$ that there exists a nonzero $\mathbf{B}_2 \in \mathcal{N}$ with $\pi(\mathbf{B}_2) = \mathbf{0} \in \mathbb{F}_q^k$. Hence $V_m(\mathbf{B}_2) \leq ms - k$, and so in both cases we get the desired bound. □

Now we come to the gist of the duality theory for digital nets. Let $\mathcal{P}$ be a digital net over $\mathbb{F}_q$ with $m \times m$ generating matrices $C^{(1)}, \ldots, C^{(s)}$ over $\mathbb{F}_q$. We set up an $m \times ms$ matrix M over $\mathbb{F}_q$ which depends only on $C^{(1)}, \ldots, C^{(s)}$ by proceeding as follows: for $j = 1, \ldots, m$, the jth row of M is obtained by concatenating the transposes of the jth columns of $C^{(1)}, \ldots, C^{(s)}$. Equivalently, the transpose $M^\top$ of M is the $ms \times m$ matrix over $\mathbb{F}_q$ that is produced by putting the $m \times m$ matrices $C^{(1)}, \ldots, C^{(s)}$ on top of each other.

Example 4.4.30 Let $q = 3$, $m = 3$, and $s = 2$, and let $\mathcal{P}$ be the digital net over $\mathbb{F}_3$ with generating matrices

$$C^{(1)} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}, \quad C^{(2)} = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

It is easily checked that $\varrho(C^{(1)}, C^{(2)}) = 3$, and so $\mathcal{P}$ is a digital $(0, 3, 2)$-net over $\mathbb{F}_3$ by Corollary 4.4.21. The matrix M associated with $\mathcal{P}$ is

$$M = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 2 & 1 & 0 \\ 1 & 2 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

Note that the transpose $M^\top$ of M can be formally written as

$$M^\top = \begin{pmatrix} C^{(1)} \\ C^{(2)} \end{pmatrix}.$$

We continue with the $m \times ms$ matrix M over $\mathbb{F}_q$ and define the *row space* of the digital net $\mathcal{P}$ over $\mathbb{F}_q$ to be the linear subspace $\mathcal{M}$ of $\mathbb{F}_q^{ms}$ generated by the row vectors of M; that is, $\mathcal{M}$ consists of all linear combinations over $\mathbb{F}_q$ of the row vectors of M. Next we take the row space $\mathcal{M}$ and form its dual space $\mathcal{M}^{\perp}$, which according to Definition 3.2.30 is given by

$$\mathcal{M}^{\perp} = \{\mathbf{B} \in \mathbb{F}_q^{ms} : \mathbf{B} \cdot \mathbf{M} = 0 \ \text{ for all } \mathbf{M} \in \mathcal{M}\},$$

where $\cdot$ denotes the dot product on $\mathbb{F}_q^{ms}$ (see Definition 3.2.17). It is obvious that $\dim(\mathcal{M}) \leq m$, and so $\mathcal{M}^{\perp}$ is a linear subspace of $\mathbb{F}_q^{ms}$ with $\dim(\mathcal{M}^{\perp}) = ms - \dim(\mathcal{M}) \geq ms - m$ by Theorem 3.2.34 (the excluded cases $k = 0$ and $k = n$ in Theorem 3.2.34 are trivial). The case of the dimension $s = 1$ is not of interest in the theory of digital nets over $\mathbb{F}_q$ since we know from Example 4.4.16 that for every integer $m \geq 1$ there exists a digital $(t, m, 1)$-net over $\mathbb{F}_q$ with the optimal value $t = 0$ of the quality parameter. Hence we can focus on the case $s \geq 2$, and then $\mathcal{M}^{\perp}$ is a nonzero linear subspace of $\mathbb{F}_q^{ms}$ since $\dim(\mathcal{M}^{\perp}) \geq ms - m \geq m \geq 1$. We can therefore talk about the minimum distance $\delta_m(\mathcal{M}^{\perp})$. The following theorem is the keystone of the duality theory for digital nets.

Theorem 4.4.31 *Let q be a prime power and let $m \geq 1$ and $s \geq 2$ be integers. Let $\mathcal{P}$ be an s-dimensional digital net over $\mathbb{F}_q$ with $m \times m$ generating matrices over $\mathbb{F}_q$ and let $\mathcal{M} \subseteq \mathbb{F}_q^{ms}$ be its row space. Then an integer t with $0 \leq t \leq m$ is a quality parameter of $\mathcal{P}$ if and only if the dual space $\mathcal{M}^{\perp}$ of $\mathcal{M}$ satisfies*

$$\delta_m(\mathcal{M}^{\perp}) \geq m - t + 1.$$

Proof. Let $C^{(1)}, \ldots, C^{(s)}$ be the generating matrices of $\mathcal{P}$. As usual, we write $\mathbf{c}_j^{(i)}$ for the jth row vector of $C^{(i)}$, where $1 \leq j \leq m$ and $1 \leq i \leq s$. In view of Theorem 4.4.19, we have to prove that the system $\{\mathbf{c}_j^{(i)} \in \mathbb{F}_q^m : 1 \leq j \leq m,\ 1 \leq i \leq s\}$ is a (d, m, s)-system over $\mathbb{F}_q$ if and only $\mathcal{M}^{\perp}$ satisfies $\delta_m(\mathcal{M}^{\perp}) \geq d + 1$.

Let M be the $m \times ms$ matrix over $\mathbb{F}_q$ constructed from $C^{(1)}, \ldots, C^{(s)}$ as above. We take any vector $\mathbf{B} \in \mathbb{F}_q^{ms}$ and write it as $\mathbf{B} = (\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(s)})$ with

$$\mathbf{b}^{(i)} = (b_1^{(i)}, \ldots, b_m^{(i)}) \in \mathbb{F}_q^m \qquad \text{for } 1 \leq i \leq s.$$

Then a linear dependence relation

$$\sum_{i=1}^{s} \sum_{j=1}^{m} b_j^{(i)} \mathbf{c}_j^{(i)} = \mathbf{0} \in \mathbb{F}_q^m \tag{4.50}$$

can be put in the form $\mathbf{B}M^{\top} = \mathbf{0} \in \mathbb{F}_q^m$ (recall the description of $M^{\top}$ prior to Example 4.4.30). Furthermore, the identity $\mathbf{B}M^{\top} = \mathbf{0}$ holds if and only if $\mathbf{B}$ is orthogonal to each column vector of $M^{\top}$ (or in other words to each row vector of M), that is, if and only if $\mathbf{B} \in \mathcal{M}^{\perp}$.

Now assume that $\{\mathbf{c}_j^{(i)} \in \mathbb{F}_q^m : 1 \leq j \leq m,\ 1 \leq i \leq s\}$ is a (d, m, s)-system over $\mathbb{F}_q$. Consider any nonzero vector $\mathbf{B} = (\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(s)}) \in \mathcal{M}^{\perp}$. Then from the above we get the linear dependence relation (4.50). Put $v_m(\mathbf{b}^{(i)}) = e_i$ for $1 \le i \le s$. Then

$$\sum_{i=1}^{s}\sum_{j=1}^{e_i} b_j^{(i)}\mathbf{c}_j^{(i)} = \mathbf{0} \in \mathbb{F}_q^m.$$

Since not all coefficients $b_j^{(i)}$ are 0, the system $\{\mathbf{c}_j^{(i)} \in \mathbb{F}_q^m : 1 \le j \le e_i,\ 1 \le i \le s\}$ is linearly dependent over $\mathbb{F}_q$. Thus, the definition of a (d, m, s)-system over $\mathbb{F}_q$ implies that $\sum_{i=1}^{s} e_i \ge d + 1$. Therefore

$$V_m(\mathbf{B}) = \sum_{i=1}^{s} v_m(\mathbf{b}^{(i)}) = \sum_{i=1}^{s} e_i \ge d + 1,$$

and so $\delta_m(\mathcal{M}^\perp) \ge d + 1$.

Conversely, assume that $\delta_m(\mathcal{M}^\perp) \ge d + 1$. We have to verify that every system $\{\mathbf{c}_j^{(i)} \in \mathbb{F}_q^m : 1 \le j \le d_i,\ 1 \le i \le s\}$ with nonnegative integers $d_1, \dots, d_s$ satisfying $\sum_{i=1}^{s} d_i = d$ is linearly independent over $\mathbb{F}_q$. Suppose, on the contrary, that such a system were linearly dependent over $\mathbb{F}_q$, that is, that there exist coefficients $b_j^{(i)} \in \mathbb{F}_q$, not all 0, such that

$$\sum_{i=1}^{s}\sum_{j=1}^{d_i} b_j^{(i)}\mathbf{c}_j^{(i)} = \mathbf{0} \in \mathbb{F}_q^m.$$

Define $b_j^{(i)} = 0$ for $d_i < j \le m$, $1 \le i \le s$. Then

$$\sum_{i=1}^{s}\sum_{j=1}^{m} b_j^{(i)}\mathbf{c}_j^{(i)} = \mathbf{0} \in \mathbb{F}_q^m.$$

This is a linear dependence relation of the form (4.50), and by what was demonstrated earlier in the proof, this leads to a nonzero vector $\mathbf{B} \in \mathcal{M}^\perp$. Hence $\delta_m(\mathcal{M}^\perp) \ge d + 1$ implies that $V_m(\mathbf{B}) \ge d + 1$. On the other hand, $v_m(\mathbf{b}^{(i)}) \le d_i$ for $1 \le i \le s$ by the definition of the $b_j^{(i)}$, and so

$$V_m(\mathbf{B}) = \sum_{i=1}^{s} v_m(\mathbf{b}^{(i)}) \le \sum_{i=1}^{s} d_i = d.$$

This is the desired contradiction. □

Corollary 4.4.32 *Let q be a prime power and let $m \ge 1$ and $s \ge 2$ be integers. Then from every linear subspace $\mathcal{N}$ of $\mathbb{F}_q^{ms}$ with $\dim(\mathcal{N}) \ge ms - m$ we can construct a digital (t, m, s)-net over $\mathbb{F}_q$ with*

$$t = m - \delta_m(\mathcal{N}) + 1.$$

Proof. Put $\mathcal{M} = \mathcal{N}^\perp \subseteq \mathbb{F}_q^{ms}$. Then $\dim(\mathcal{M}) = ms - \dim(\mathcal{N}) \leq m$. By using basis vectors of $\mathcal{M}$ as row vectors and supplementing them by zero vectors from $\mathbb{F}_q^{ms}$ as needed, we can set up an $m \times ms$ matrix M over $\mathbb{F}_q$ whose row vectors generate $\mathcal{M}$. In the same way as $m \times m$ generating matrices $C^{(1)}, \ldots, C^{(s)}$ over $\mathbb{F}_q$ led to an $m \times ms$ matrix over $\mathbb{F}_q$, we can start from the matrix M and recover $m \times m$ matrices $C^{(1)}, \ldots, C^{(s)}$ over $\mathbb{F}_q$. Let $\mathcal{P}$ be the digital net over $\mathbb{F}_q$ with generating matrices $C^{(1)}, \ldots, C^{(s)}$. Then by construction, $\mathcal{M}$ is the row space of $\mathcal{P}$. Theorem 4.4.31 shows that the possible values of the quality parameter t of $\mathcal{P}$ satisfy $t \geq m - \delta_m(\mathcal{M}^\perp) + 1 = m - \delta_m(\mathcal{N}) + 1$. The least possible value is $t = m - \delta_m(\mathcal{N}) + 1$, but we have to ascertain that it satisfies $0 \leq t \leq m$. The inequality $t \leq m$ is trivial since $\delta_m(\mathcal{N}) \geq 1$. Furthermore, Proposition 4.4.29 implies that

$$t = m - \delta_m(\mathcal{N}) + 1 \geq m - ms + \dim(\mathcal{N}) \geq 0$$

since $\dim(\mathcal{N}) \geq ms - m$ by assumption. □

Example 4.4.33 We return to the digital $(0, m, 2)$-net $\mathcal{P}$ over $\mathbb{F}_q$ in Example 4.4.17 and discuss it from the viewpoint of duality theory. Since the generating matrices $C^{(1)}$ and $C^{(2)}$ of $\mathcal{P}$ are symmetric, we can write the $m \times 2m$ matrix M over $\mathbb{F}_q$ in an obvious notation as $M = \left(C^{(1)} \mid C^{(2)}\right)$. Therefore the row space $\mathcal{M}$ of $\mathcal{P}$ consists of all vectors

$$\mathbf{C} = (c_1, c_2, \ldots, c_m, c_m, \ldots, c_2, c_1) \in \mathbb{F}_q^{2m} \tag{4.51}$$

with $c_1, \ldots, c_m$ running independently through $\mathbb{F}_q$. Now we take a vector $\mathbf{B}$ of the form

$$\mathbf{B} = (b_1, b_2, \ldots, b_m, -b_m, \ldots, -b_2, -b_1) \in \mathbb{F}_q^{2m} \tag{4.52}$$

with arbitrary $b_1, \ldots, b_m \in \mathbb{F}_q$. Then for every vector $\mathbf{C}$ in (4.51) we get $\mathbf{B} \cdot \mathbf{C} = 0$, and so $\mathbf{B} \in \mathcal{M}^\perp$. Note that $\dim(\mathcal{M}^\perp) = 2m - \dim(\mathcal{M}) = m$, and so the vectors $\mathbf{B}$ in (4.52) yield exactly the dual space $\mathcal{M}^\perp$ of $\mathcal{M}$. We want to deduce from Theorem 4.4.31 that $t = 0$ is a quality parameter of the digital net $\mathcal{P}$ over $\mathbb{F}_q$. To this end, we have to prove that $\delta_m(\mathcal{M}^\perp) \geq m + 1$. Consequently, we take any nonzero vector $\mathbf{B}$ in (4.52) and we write as usual $\mathbf{B} = (\mathbf{b}^{(1)}, \mathbf{b}^{(2)})$, here with $\mathbf{b}^{(1)} = (b_1, b_2, \ldots, b_m) \in \mathbb{F}_q^m$ and $\mathbf{b}^{(2)} = (-b_m, \ldots, -b_2, -b_1) \in \mathbb{F}_q^m$. We must have $\mathbf{b}^{(1)} \neq \mathbf{0} \in \mathbb{F}_q^m$, so let us say that $v_m(\mathbf{b}^{(1)}) = j \geq 1$. Then $b_j \neq 0$, so in the vector $\mathbf{b}^{(2)}$ we find the coordinate $-b_j \neq 0$ in the position $m - j + 1$. Therefore $v_m(\mathbf{b}^{(2)}) \geq m - j + 1$, and so $V_m(\mathbf{B}) = v_m(\mathbf{b}^{(1)}) + v_m(\mathbf{b}^{(2)}) \geq j + m - j + 1 = m + 1$. Thus, we get indeed $\delta_m(\mathcal{M}^\perp) \geq m + 1$.

Example 4.4.34 Consider the digital $(0, 3, 2)$-net $\mathcal{P}$ over $\mathbb{F}_3$ in Example 4.4.30. The matrix M associated with $\mathcal{P}$ is given in that example, and so the row space $\mathcal{M}$

of $\mathcal{P}$ consists of all linear combinations over $\mathbb{F}_3$ of the row vectors of M. The vectors

$$\begin{aligned}\mathbf{B}_1 &= (2,2,2,1,0,0) \in \mathbb{F}_3^6,\\ \mathbf{B}_2 &= (0,2,1,0,1,0) \in \mathbb{F}_3^6,\\ \mathbf{B}_3 &= (0,0,2,0,0,1) \in \mathbb{F}_3^6\end{aligned}$$

are orthogonal to all row vectors of M, and so $\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3 \in \mathcal{M}^\perp$. Since $\dim(\mathcal{M}^\perp) = 6 - \dim(\mathcal{M}) = 3$ and $\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3$ are linearly independent over $\mathbb{F}_3$, the subspace $\mathcal{M}^\perp$ of $\mathbb{F}_3^6$ consists of all linear combinations over $\mathbb{F}_3$ of $\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3$. We want to show by Theorem 4.4.31 that $t = 0$ is a quality parameter of the digital net $\mathcal{P}$ over $\mathbb{F}_3$. This will be achieved if we prove that $\delta_3(\mathcal{M}^\perp) \geq 4$. We take any nonzero vector $\mathbf{B} \in \mathcal{M}^\perp$ and write as usual $\mathbf{B} = (\mathbf{b}^{(1)}, \mathbf{b}^{(2)})$ with $\mathbf{b}^{(1)} \in \mathbb{F}_3^3$ and $\mathbf{b}^{(2)} \in \mathbb{F}_3^3$. Note that for every nontrivial linear combination over $\mathbb{F}_3$ of $\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3$, we obtain $v_3(\mathbf{b}^{(1)}) \geq 1$ and $v_3(\mathbf{b}^{(2)}) \geq 1$. If $V_3(\mathbf{B}) = 2$, then necessarily $v_3(\mathbf{b}^{(2)}) = 1$. But then $\mathbf{B} = a\mathbf{B}_1$ for some $a \in \mathbb{F}_3^*$, hence $V_3(\mathbf{B}) = 4$, a contradiction. If $V_3(\mathbf{B}) = 3$, then again we cannot have $v_3(\mathbf{b}^{(2)}) = 1$, and so we must have $v_3(\mathbf{b}^{(2)}) = 2$ and $v_3(\mathbf{b}^{(1)}) = 1$. We deduce from $v_3(\mathbf{b}^{(2)}) = 2$ that $\mathbf{B} = b\mathbf{B}_1 + c\mathbf{B}_2$ with $b, c \in \mathbb{F}_3$ and $c \neq 0$. But then $v_3(\mathbf{b}^{(1)}) \geq 2$ and $V_3(\mathbf{B}) \geq 4$, again a contradiction. Thus, we must have $\delta_3(\mathcal{M}^\perp) \geq 4$, and since $V_3(\mathbf{B}_1) = 4$, we get in fact $\delta_3(\mathcal{M}^\perp) = 4$.

4.4.3 Constructions of digital nets

We have already seen some examples of digital nets in the preceding subsection, and now we present several systematic constructions of larger families of digital nets. We start with an intriguing application of linear codes to the theory of digital nets. The broad impact of coding theory on digital nets is in fact a remarkable phenomenon, and it can be expected that more links between these areas will be discovered in the future. We use the standard notation for linear codes from Section 3.2, namely that a linear $[n, k, d]$ code over $\mathbb{F}_q$ is a linear code over $\mathbb{F}_q$ of length n, dimension k, and minimum distance d.

Theorem 4.4.35 *Let q be a prime power and let n, k, and d be integers with $3 \leq d \leq n$ and $1 \leq k \leq n-1$. Then from a linear $[n, k, d]$ code over $\mathbb{F}_q$ we can derive a digital $(n-k-d+1, n-k, s)$-net over $\mathbb{F}_q$, where $s = \lfloor 2n/(d-1) \rfloor$ if d is odd and $s = \lfloor (2n-2)/(d-2) \rfloor$ if d is even.*

Proof. We first note that we always have $n - k - d + 1 \geq 0$ by the Singleton bound for linear codes in Corollary 3.4.11. Furthermore, the definition of s implies that $s \geq 2$.

Let H be a parity-check matrix of a given linear $[n, k, d]$ code over $\mathbb{F}_q$. Then H is an $(n-k) \times n$ matrix over $\mathbb{F}_q$ whose column vectors $\mathbf{h}_1, \ldots, \mathbf{h}_n \in \mathbb{F}_q^{n-k}$ have the property that any $d-1$ of them are linearly independent over $\mathbb{F}_q$ (see Theorem

3.2.44). In view of Theorem 4.4.19, it suffices to derive a $(d-1, n-k, s)$-system $\{\mathbf{c}_j^{(i)} \in \mathbb{F}_q^{n-k} : 1 \le j \le n-k,\ 1 \le i \le s\}$ over $\mathbb{F}_q$.

We commence with the case where $d \ge 3$ is odd and we put $a = (d-1)/2$. Note that then $s = \lfloor n/a \rfloor$. Now we determine the vectors $\mathbf{c}_j^{(i)} \in \mathbb{F}_q^{n-k}$, $1 \le j \le 2a = d-1$, $1 \le i \le s$, according to the following table.

	$i=1$	$i=2$	$i=3$	$\dots$	$i=s$
$\mathbf{c}_1^{(i)}$	$\mathbf{h}_1$	$\mathbf{h}_{a+1}$	$\mathbf{h}_{2a+1}$	$\dots$	$\mathbf{h}_{(s-1)a+1}$
$\mathbf{c}_2^{(i)}$	$\mathbf{h}_2$	$\mathbf{h}_{a+2}$	$\mathbf{h}_{2a+2}$	$\dots$	$\mathbf{h}_{(s-1)a+2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\mathbf{c}_a^{(i)}$	$\mathbf{h}_a$	$\mathbf{h}_{2a}$	$\mathbf{h}_{3a}$	$\dots$	$\mathbf{h}_{sa}$
$\mathbf{c}_{a+1}^{(i)}$	$\mathbf{h}_{2a}$	$\mathbf{h}_a$	$\mathbf{h}_a$	$\dots$	$\mathbf{h}_a$
$\mathbf{c}_{a+2}^{(i)}$	$\mathbf{h}_{2a-1}$	$\mathbf{h}_{a-1}$	$\mathbf{h}_{a-1}$	$\dots$	$\mathbf{h}_{a-1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\mathbf{c}_{2a}^{(i)}$	$\mathbf{h}_{a+1}$	$\mathbf{h}_1$	$\mathbf{h}_1$	$\dots$	$\mathbf{h}_1$

The vectors $\mathbf{c}_j^{(i)}$ for $2a+1 = d \le j \le n-k$ and $1 \le i \le s$ can be chosen arbitrarily in $\mathbb{F}_q^{n-k}$. The largest subscript r of a vector $\mathbf{h}_r$ in the table above is $r = sa = \lfloor n/a \rfloor a \le n$, and so all entries in the table make sense.

We claim that $\{\mathbf{c}_j^{(i)} \in \mathbb{F}_q^{n-k} : 1 \le j \le n-k,\ 1 \le i \le s\}$ is a $(d-1, n-k, s)$-system over $\mathbb{F}_q$. For any nonnegative integers $d_1, \dots, d_s$ with $\sum_{i=1}^s d_i = d-1 = 2a$, we have to show that the system

$$S = S_{d_1,\dots,d_s} = \{\mathbf{c}_j^{(i)} \in \mathbb{F}_q^{n-k} : 1 \le j \le d_i,\ 1 \le i \le s\}$$

is linearly independent over $\mathbb{F}_q$. Note first that all vectors $\mathbf{h}_r$ in the upper $a \times s$ subarray of the table above have different subscripts r, and so any $d-1$ of them are linearly independent over $\mathbb{F}_q$ by the given linear independence property of the vectors $\mathbf{h}_1, \dots, \mathbf{h}_n$. Consequently, if $d_i \le a$ for $1 \le i \le s$, then the system S is linearly independent over $\mathbb{F}_q$. In the remaining case, $d_i > a$ holds for some i, and since $\sum_{i=1}^s d_i = 2a$, this holds for exactly one $i = i_0$. If $i_0 = 1$, then $d_1 = a+b$ with $1 \le b \le a$, and from the column $i=1$ in the table above we pick the vectors

$$\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_a, \mathbf{h}_{2a}, \mathbf{h}_{2a-1}, \dots, \mathbf{h}_{2a-b+1}.$$

From the other columns we select altogether $a-b$ vectors, and in the entire collection of the chosen vectors $\mathbf{h}_r$ all subscripts r are different. Hence again the corresponding system S is linearly independent over $\mathbb{F}_q$. Finally, if $i_0 \ge 2$, then $d_{i_0} = a+b$ with $1 \le b \le a$, and from the column $i = i_0$ in the table above we pick the vectors

$$\mathbf{h}_{(i_0-1)a+1}, \mathbf{h}_{(i_0-1)a+2}, \dots, \mathbf{h}_{i_0 a}, \mathbf{h}_a, \mathbf{h}_{a-1}, \dots, \mathbf{h}_{a-b+1}.$$

From the other columns we select altogether $a-b$ vectors, and for the same reason as before the corresponding system S is linearly independent over $\mathbb{F}_q$. This completes the proof for the case where d is odd.

Now we consider the case where $d \geq 4$ is even and we put $a = (d-2)/2$. Note that then $s = \lfloor (n-1)/a \rfloor$. We determine the vectors $\mathbf{c}_j^{(i)} \in \mathbb{F}_q^{n-k}$, $1 \leq j \leq 2a+1 = d-1$, $1 \leq i \leq s$, according to the following table.

	$i=1$	$i=2$	$i=3$	$\dots$	$i=s$
$\mathbf{c}_1^{(i)}$	$\mathbf{h}_1$	$\mathbf{h}_{a+1}$	$\mathbf{h}_{2a+1}$	$\dots$	$\mathbf{h}_{(s-1)a+1}$
$\mathbf{c}_2^{(i)}$	$\mathbf{h}_2$	$\mathbf{h}_{a+2}$	$\mathbf{h}_{2a+2}$	$\dots$	$\mathbf{h}_{(s-1)a+2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\mathbf{c}_a^{(i)}$	$\mathbf{h}_a$	$\mathbf{h}_{2a}$	$\mathbf{h}_{3a}$	$\dots$	$\mathbf{h}_{sa}$
$\mathbf{c}_{a+1}^{(i)}$	$\mathbf{h}_{sa+1}$	$\mathbf{h}_{sa+1}$	$\mathbf{h}_{sa+1}$	$\dots$	$\mathbf{h}_{sa+1}$
$\mathbf{c}_{a+2}^{(i)}$	$\mathbf{h}_{2a}$	$\mathbf{h}_a$	$\mathbf{h}_a$	$\dots$	$\mathbf{h}_a$
$\mathbf{c}_{a+3}^{(i)}$	$\mathbf{h}_{2a-1}$	$\mathbf{h}_{a-1}$	$\mathbf{h}_{a-1}$	$\dots$	$\mathbf{h}_{a-1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\mathbf{c}_{2a+1}^{(i)}$	$\mathbf{h}_{a+1}$	$\mathbf{h}_1$	$\mathbf{h}_1$	$\dots$	$\mathbf{h}_1$

The vectors $\mathbf{c}_j^{(i)}$ for $2a+2 = d \leq j \leq n-k$ and $1 \leq i \leq s$ can be chosen arbitrarily from $\mathbb{F}_q^{n-k}$. The largest subscript r of a vector $\mathbf{h}_r$ in the table above is $r = sa+1 = \lfloor (n-1)/a \rfloor a + 1 \leq n$, and so all entries in the table make sense.

We claim that $\{\mathbf{c}_j^{(i)} \in \mathbb{F}_q^{n-k} : 1 \leq j \leq n-k,\ 1 \leq i \leq s\}$ is a $(d-1, n-k, s)$-system over $\mathbb{F}_q$. For any nonnegative integers $d_1, \ldots, d_s$ with $\sum_{i=1}^s d_i = d-1 = 2a+1$, we have to verify that the system

$$T = T_{d_1,\ldots,d_s} = \{\mathbf{c}_j^{(i)} \in \mathbb{F}_q^{n-k} : 1 \leq j \leq d_i,\ 1 \leq i \leq s\}$$

is linearly independent over $\mathbb{F}_q$. As before, the case where $d_i \leq a$ for $1 \leq i \leq s$ is obvious. In the remaining case, there is again exactly one $i = i_0$ with $d_i > a$. A similar analysis as in the case d odd shows that the system T is linearly independent over $\mathbb{F}_q$. The only new potential problem is that we may pick the vector $\mathbf{h}_{sa+1}$ twice, but since this vector appears only in row $a+1$ of the table above and since $\sum_{i=1}^s d_i = 2a+1 < 2(a+1)$, this cannot happen. □

Example 4.4.36 If we want to achieve the optimal value $t = 0$ of the quality parameter in Theorem 4.4.35, then we have to use a linear $[n, k, d]$ code over $\mathbb{F}_q$ with $d = n-k+1$, that is, an MDS code over $\mathbb{F}_q$ (see Definition 3.4.13). An interesting family of MDS codes is formed by the generalized Reed-Solomon codes in Remark 3.5.34. In particular, for every prime power q and every integer k with $1 \leq k \leq q$, we obtain a linear $[q, k, q-k+1]$ code over $\mathbb{F}_q$ from Remark 3.5.34. We first choose $q \geq 3$ and $k = q-2$. Then $d = q-k+1 = 3$ is odd, and so $s = \lfloor 2q/2 \rfloor = q$ in Theorem 4.4.35. Therefore Theorem 4.4.35 yields a digital $(0, 2, q)$-net over $\mathbb{F}_q$. Next we choose $q \geq 4$ and $k = q-3$. Then $d = q-k+1 = 4$ is even, and so $s = \lfloor (2q-2)/2 \rfloor = q-1$ in Theorem 4.4.35. Therefore Theorem 4.4.35 yields a digital $(0, 3, q-1)$-net over $\mathbb{F}_q$. By similar arguments, the choice $q \geq 5$ and $k = q-4$ yields a digital $(0, 4, \lfloor q/2 \rfloor)$-net over $\mathbb{F}_q$, and many other examples of this type can be derived from Theorem 4.4.35 and generalized Reed-Solomon codes.

A rather wide family of digital nets is that of hyperplane nets introduced by Pirsic, Dick, and Pillichshammer [161]. The construction of hyperplane nets is based on the duality theory for digital nets described in Subsection 4.4.2. Let q be a prime power and let $m \geq 1$ and $s \geq 2$ be integers. Together with the finite field $\mathbb{F}_q$ we consider also its extension field $\mathbb{F}_{q^m}$ with q^m elements. Let $W = \mathbb{F}_{q^m}^s$ be the standard s-dimensional vector space over $\mathbb{F}_{q^m}$. Then W can be viewed also as a vector space over $\mathbb{F}_q$, and since W contains exactly $(q^m)^s = q^{ms}$ vectors, it follows from Proposition 3.2.6 that W has dimension ms as a vector space over $\mathbb{F}_q$. Next we choose an ordered basis B of $\mathbb{F}_{q^m}$ over $\mathbb{F}_q$, for instance as in Remark 3.2.7. Now let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_s) \in W$ be arbitrary. For each $i = 1, \ldots, s$, the element $\beta_i \in \mathbb{F}_{q^m}$ has a coordinate vector $\tau(\beta_i) \in \mathbb{F}_q^m$ relative to the ordered basis B. Then

$$\sigma(\boldsymbol{\beta}) = (\tau(\beta_1), \ldots, \tau(\beta_s)) \in \mathbb{F}_q^{ms} \qquad \text{for all } \boldsymbol{\beta} \in W$$

defines a bijective linear transformation $\sigma : W \to \mathbb{F}_q^{ms}$ from W onto $\mathbb{F}_q^{ms}$.

For the construction of a hyperplane net, we fix a vector $\boldsymbol{\alpha} \in W$ and we put

$$W_{\boldsymbol{\alpha}} := \{\boldsymbol{\beta} \in W : \boldsymbol{\alpha} \cdot \boldsymbol{\beta} = 0\},$$

where $\cdot$ is the dot product on W. It is clear that $W_{\boldsymbol{\alpha}}$ is a linear subspace of W, both for W as a vector space over $\mathbb{F}_{q^m}$ and over $\mathbb{F}_q$. If $\boldsymbol{\alpha} = \mathbf{0} \in W$, then $W_{\boldsymbol{\alpha}} = W$, and so $\dim(W_{\boldsymbol{\alpha}}) = ms$ as a vector space over $\mathbb{F}_q$. If $\boldsymbol{\alpha} \neq \mathbf{0}$, then $W_{\boldsymbol{\alpha}}$ contains exactly $(q^m)^{s-1} = q^{ms-m}$ vectors, and so $\dim(W_{\boldsymbol{\alpha}}) = ms - m$ as a vector space over $\mathbb{F}_q$. Note that if $\boldsymbol{\alpha} \neq \mathbf{0}$, then $W_{\boldsymbol{\alpha}}$ may be interpreted geometrically as a hyperplane, and this explains the terminology "hyperplane net". In any case, we can say that $\dim(W_{\boldsymbol{\alpha}}) \geq ms - m$. Next we let $\mathcal{N}_{\boldsymbol{\alpha}} = \sigma(W_{\boldsymbol{\alpha}}) \subseteq \mathbb{F}_q^{ms}$ be the image of $W_{\boldsymbol{\alpha}}$ under σ. Since σ is a bijective linear transformation, $\mathcal{N}_{\boldsymbol{\alpha}}$ is a linear subspace of $\mathbb{F}_q^{ms}$ with $\dim(\mathcal{N}_{\boldsymbol{\alpha}}) = \dim(W_{\boldsymbol{\alpha}}) \geq ms - m$, and so we are in a position to apply Corollary 4.4.32.

Definition 4.4.37 Let q be a prime power and let $m \geq 1$ and $s \geq 2$ be integers. For a given $\boldsymbol{\alpha} \in W = \mathbb{F}_{q^m}^s$, the digital net $\mathcal{P}_{\boldsymbol{\alpha}}$ over $\mathbb{F}_q$ obtained from Corollary 4.4.32 by using the linear subspace $\mathcal{N}_{\boldsymbol{\alpha}}$ of $\mathbb{F}_q^{ms}$ constructed above is called a *hyperplane net* over $\mathbb{F}_q$.

It follows from Corollary 4.4.32 that $\mathcal{P}_{\boldsymbol{\alpha}}$ is a digital (t, m, s)-net over $\mathbb{F}_q$ with $t = m - \delta_m(\mathcal{N}_{\boldsymbol{\alpha}}) + 1$. As usual in the duality theory for digital nets, we strive to make the minimum distance $\delta_m(\mathcal{N}_{\boldsymbol{\alpha}})$ as large as possible. The following result establishes a lower bound on $\delta_m(\mathcal{N}_{\boldsymbol{\alpha}})$ that can always be satisfied by a suitable choice of $\boldsymbol{\alpha} \in W$.

Theorem 4.4.38 *For every prime power q and all integers $m \geq 1$ and $s \geq 2$, there exists an $\boldsymbol{\alpha} \in W = \mathbb{F}_{q^m}^s$ such that*

$$\delta_m(\mathcal{N}_{\boldsymbol{\alpha}}) \geq m + 1 - \lceil (s-1) \log_q(m+1) \rceil,$$

where $\log_q$ denotes the logarithm to the base q.

Proof. We proceed by an elimination method: we weed out those $\boldsymbol{\alpha} \in W$ for which $\delta_m(\mathcal{N}_{\boldsymbol{\alpha}}) \leq m - \lceil (s-1) \log_q(m+1) \rceil$ and we show that there is still an $\boldsymbol{\alpha}_0 \in W$ left over, which then necessarily satisfies $\delta_m(\mathcal{N}_{\boldsymbol{\alpha}_0}) \geq m + 1 - \lceil (s-1) \log_q(m+1) \rceil$.

For every integer d with $0 \leq d \leq m$, the number of $\mathbf{b} \in \mathbb{F}_q^m$ with $v_m(\mathbf{b}) = d$ is given by $\varepsilon_d q^d$, where $\varepsilon_0 = 1$ and $\varepsilon_d = (q-1)/q$ for $1 \leq d \leq m$. Thus, for a fixed integer k with $1 \leq k \leq ms$ and for $(d_1, \ldots, d_s) \in \mathbb{Z}^s$ with $0 \leq d_i \leq m$ for $1 \leq i \leq s$ and $\sum_{i=1}^s d_i = k$, the number of $\mathbf{B} = (\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(s)}) \in \mathbb{F}_q^{ms}$ with $v_m(\mathbf{b}^{(i)}) = d_i$ for $1 \leq i \leq s$ is equal to $\prod_{i=1}^s \varepsilon_{d_i} q^{d_i}$. This number is at most

$$(q-1) q^{d_1 + \cdots + d_s - 1} = (q-1) q^{k-1}$$

since at least one d_i is positive. Furthermore, the number of $(d_1, \ldots, d_s) \in \mathbb{Z}^s$ with $0 \leq d_i \leq m$ for $1 \leq i \leq s$ and $\sum_{i=1}^s d_i = k$ is bounded from above by $(m+1)^{s-1}$, since for each of $d_1, \ldots, d_{s-1}$ there are at most $m+1$ possibilities and there is at most one choice for d_s for any given $d_1, \ldots, d_{s-1}$, and k. Altogether, we have shown that the number $A_q(k, m, s)$ of $\mathbf{B} \in \mathbb{F}_q^{ms}$ with $V_m(\mathbf{B}) = k$ satisfies

$$A_q(k, m, s) \leq (q-1)(m+1)^{s-1} q^{k-1}.$$

Next we estimate the number of $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_s) \in W$ with $\delta_m(\mathcal{N}_{\boldsymbol{\alpha}}) = k$. For such an $\boldsymbol{\alpha}$, there exists a vector $\mathbf{B} \in \mathcal{N}_{\boldsymbol{\alpha}}$ with $V_m(\mathbf{B}) = k$. We take such a $\mathbf{B} \in \mathcal{N}_{\boldsymbol{\alpha}}$ and note that by the definition of $\mathcal{N}_{\boldsymbol{\alpha}}$ there exists a unique $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_s) \in W_{\boldsymbol{\alpha}} \setminus \{\mathbf{0}\}$ with $\sigma(\boldsymbol{\beta}) = \mathbf{B}$. Now $\boldsymbol{\beta} \in W_{\boldsymbol{\alpha}}$ means that

$$\boldsymbol{\alpha} \cdot \boldsymbol{\beta} = \alpha_1 \beta_1 + \cdots + \alpha_s \beta_s = 0.$$

Since at least one $\beta_i \neq 0$, the number of $\boldsymbol{\alpha} \in W$ with $\boldsymbol{\alpha} \cdot \boldsymbol{\beta} = 0$ is exactly $q^{m(s-1)}$. It follows that the number of $\boldsymbol{\alpha} \in W$ with $\delta_m(\mathcal{N}_{\boldsymbol{\alpha}}) = k$ is at most

$$A_q(k, m, s) q^{m(s-1)} \leq (q-1)(m+1)^{s-1} q^{m(s-1)+k-1}.$$

Finally, we put $K = m - \lceil (s-1) \log_q(m+1) \rceil$ and we note that we can assume $K \geq 1$, for otherwise the theorem is trivial. Then by what we have already shown, the number of $\boldsymbol{\alpha} \in W$ with $\delta_m(\mathcal{N}_{\boldsymbol{\alpha}}) \leq K$ is at most

$$\begin{aligned}(q-1)(m+1)^{s-1} q^{m(s-1)} \sum_{k=1}^{K} q^{k-1} &< (m+1)^{s-1} q^{m(s-1)+K} \\ &\leq (m+1)^{s-1} q^{m(s-1)+m-(s-1)\log_q(m+1)} = q^{ms}.\end{aligned}$$

The set W has q^{ms} elements, and so there exists an $\boldsymbol{\alpha}_0 \in W$ with $\delta_m(\mathcal{N}_{\boldsymbol{\alpha}_0}) \geq K+1$. □

Corollary 4.4.39 *For every prime power q and all integers $m \geq 1$ and $s \geq 2$, there exists an $\boldsymbol{\alpha} \in W = \mathbb{F}_{q^m}^s$ such that the hyperplane net $\mathcal{P}_{\boldsymbol{\alpha}}$ in Definition 4.4.37 is a digital (t, m, s)-net over $\mathbb{F}_q$ with*

$$t \leq \lceil (s-1) \log_q (m+1) \rceil.$$

Proof. This follows from Corollary 4.4.32 and Theorem 4.4.38. □

If we choose an $\boldsymbol{\alpha} \in W$ according to Corollary 4.4.39, then it follows from Theorem 4.4.14 that the star discrepancy $D_N^*(\mathcal{P}_{\boldsymbol{\alpha}})$ of the corresponding hyperplane net $\mathcal{P}_{\boldsymbol{\alpha}}$ with $N = q^m$ satisfies

$$D_N^*(\mathcal{P}_{\boldsymbol{\alpha}}) \leq c(q,s) N^{-1} (\log N)^{2s-2}$$

with a constant $c(q,s) > 0$ depending only on q and s. It turns out that for $s \geq 3$ one can find even better hyperplane nets by using the general principle of the CBC algorithm for good lattice points (see Algorithm 4.3.12). In the present context, the idea is to construct a good vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_s) \in W$ coordinate by coordinate, by starting with $\alpha_1 = 1 \in \mathbb{F}_{q^m}$ and computing $\alpha_{d+1} \in \mathbb{F}_{q^m}$, $1 \leq d \leq s-1$, by minimizing a certain quantity that depends on the previously computed coordinates $\alpha_1, \ldots, \alpha_d$ and a variable element γ ranging over $\mathbb{F}_{q^m}$. This has the effect that in the bound on $D_N^*(\mathcal{P}_{\boldsymbol{\alpha}})$ above we can replace the exponent $2s-2$ of $\log N$ by s, provided that we use a vector $\boldsymbol{\alpha} \in W$ obtained from this CBC algorithm. We refer to [39, Section 11.3] for the details.

We mentioned good lattice points in the paragraph above, and there is actually an analog of lattice point sets in the context of digital nets. The construction of lattice point sets in Section 4.3 is based on the arithmetic of rational numbers, whereas the analog for digital nets employs the arithmetic of rational functions over a finite field. Let $\mathbb{F}_q(x)$ be the field of rational functions over $\mathbb{F}_q$ which consists of all fractions $g(x)/f(x)$ of polynomials with a numerator $g(x) \in \mathbb{F}_q[x]$ and a nonzero denominator $f(x) \in \mathbb{F}_q[x]$. Here q is as usual an arbitrary prime power. The arithmetic in $\mathbb{F}_q(x)$ is as expected; the only difference compared to classical rational functions, say over the real numbers, is that the arithmetic for the coefficients is performed in the finite field $\mathbb{F}_q$.

We need a technical tool, namely the expansion of a rational function over $\mathbb{F}_q$ into a sort of power series. We will require the same tool again in Subsection 4.4.5 on the construction of (t,s)-sequences. So bear with us in a brief interlude about expanding rational functions into formal Laurent series. For a rational function

$g(x)/f(x) \in \mathbb{F}_q(x)$ as above, let the leading term of $f(x)$ be ax^m with $a \in \mathbb{F}_q^*$ and an integer $m = \deg(f(x)) \geq 0$. Then we can write

$$f(x) = ax^m(1 - b_1x^{-1} - \cdots - b_mx^{-m})$$

for some $b_1, \ldots, b_m \in \mathbb{F}_q$. By proceeding in a purely formal way, we obtain

$$\begin{aligned} \frac{g(x)}{f(x)} &= \frac{g(x)}{ax^m(1 - b_1x^{-1} - \cdots - b_mx^{-m})} \\ &= a^{-1}x^{-m}g(x)\sum_{k=0}^{\infty}(b_1x^{-1} + \cdots + b_mx^{-m})^k, \end{aligned}$$

and so finally

$$\frac{g(x)}{f(x)} = \sum_{r=w}^{\infty} e_r x^{-r} \tag{4.53}$$

with coefficients $e_r \in \mathbb{F}_q$ and an integer w. The formal expression $\sum_{r=w}^{\infty} e_r x^{-r}$ is called a *formal Laurent series* over $\mathbb{F}_q$ in the variable x^{-1}. It can contain infinitely many powers $x^{-1}, x^{-2}, \ldots$ with negative exponents, but only finitely many powers of x with nonnegative exponents. In contrast to power series in real and complex analysis, there is no issue of convergence for formal Laurent series. Two formal Laurent series over $\mathbb{F}_q$ are identical if for all powers of x (with arbitrary exponents from $\mathbb{Z}$) the two corresponding coefficients agree. Formal Laurent series over $\mathbb{F}_q$ can be added and multiplied just as is done for polynomials, or as you learned to do for power series in real and complex analysis. There is also division for formal Laurent series, but we do not need this operation. Altogether, the set $\mathbb{F}_q((x^{-1}))$ of all formal Laurent series over $\mathbb{F}_q$ in the variable x^{-1} forms a field. The fact that there exists the expansion $g(x)/f(x) = \sum_{r=w}^{\infty} e_r x^{-r}$ in (4.53) can be interpreted as saying that $\mathbb{F}_q(x)$ is a subfield of $\mathbb{F}_q((x^{-1}))$. We note in passing that if $\deg(g(x)) < \deg(f(x))$, then we can take $w \geq 1$ in (4.53).

Example 4.4.40 Let $q = 3$ and consider the rational function $(x + 1)/(x^2 + 1) \in \mathbb{F}_3(x)$. By proceeding as in the computation leading to (4.53), we obtain

$$\begin{aligned} \frac{x+1}{x^2+1} &= \frac{x+1}{x^2(1-2x^{-2})} = x^{-2}(x+1)(1 + 2x^{-2} + x^{-4} + 2x^{-6} + \cdots) \\ &= (x+1)(x^{-2} + 2x^{-4} + x^{-6} + 2x^{-8} + \cdots) \\ &= x^{-1} + x^{-2} + 2x^{-3} + 2x^{-4} + x^{-5} + x^{-6} + \cdots \in \mathbb{F}_3((x^{-1})). \end{aligned}$$

Since $\deg(x + 1) < \deg(x^2 + 1)$, only powers of x with negative exponents appear in the formal Laurent series expansion of $(x + 1)/(x^2 + 1)$.

Next we introduce a degree map ν on $\mathbb{F}_q((x^{-1}))$ which is a natural extension of the degree map on the polynomial ring $\mathbb{F}_q[x]$. We put $\nu(f) = \deg(f)$ for a nonzero $f \in \mathbb{F}_q[x]$ as well as $\nu(0) = -\infty$ for the zero polynomial $0 \in \mathbb{F}_q[x]$. For a nonzero rational function $g/f \in \mathbb{F}_q(x)$ with nonzero $g, f \in \mathbb{F}_q[x]$, we put $\nu(g/f) = \nu(g) - \nu(f)$. Note that in this definition it does not matter whether g/f is in reduced form or not. Finally, if a nonzero $E = \sum_{r=w}^{\infty} e_r x^{-r} \in \mathbb{F}_q((x^{-1}))$ is given, then we can assume that $e_w \neq 0$. Then by definition $\nu(E) = -w$, that is, $\nu(E)$ is the largest exponent of x that actually appears in the expression for E. This is of course in line with the general idea of a degree. So, for instance, if $E = x^{-3} + x^{-5} + x^{-6} + \cdots \in \mathbb{F}_2((x^{-1}))$, then $\nu(E) = -3$. The computation leading to (4.53) demonstrates that if $E = g/f$ is a nonzero rational function over $\mathbb{F}_q$, then the definition of $\nu(E)$ and the earlier definition of $\nu(g/f)$ agree. We have the usual rules for degrees, such as $\nu(E_1 E_2) = \nu(E_1) + \nu(E_2)$ for all $E_1, E_2 \in \mathbb{F}_q((x^{-1}))$.

Now that we are familiar with formal Laurent series expansions of rational functions over $\mathbb{F}_q$, we employ this device in a construction of digital nets over $\mathbb{F}_q$ which is due to Niederreiter [133]. As promised, these digital nets will by and large be analogous to lattice point sets, although this is not evident at the outset. As often in the construction of digital nets, we assume that the dimension s satisfies $s \geq 2$ since the one-dimensional case is trivial (see Example 4.4.16). We choose a polynomial $f \in \mathbb{F}_q[x]$ with $\deg(f) = m \geq 1$. This polynomial plays a similar role as the modulus N in the point set $\mathcal{P}(\mathbf{g}, N)$ defined in Subsection 4.3.1. Furthermore, we choose polynomials $g_1, \ldots, g_s \in \mathbb{F}_q[x]$ with $\deg(g_i) < m$ for $1 \leq i \leq s$ and we collect them in the s-tuple $\boldsymbol{g} = (g_1, \ldots, g_s) \in \mathbb{F}_q[x]^s$. The polynomial f and the s-tuple $\boldsymbol{g}$ are the basic ingredients of the construction. It is convenient to use again a notation that we introduced in Section 3.3 on cyclic codes, namely for an integer $m \geq 1$ we write $\mathbb{F}_q[x]_{<m}$ for the set of all polynomials $g \in \mathbb{F}_q[x]$ with $\deg(g) < m$. Consequently, we will often write $\boldsymbol{g} \in \mathbb{F}_q[x]_{<m}^s$ for an s-tuple $\boldsymbol{g}$ as above.

The actual construction of the digital net proceeds by defining its s generating matrices $C^{(1)}, \ldots, C^{(s)}$ over $\mathbb{F}_q$. For each $i = 1, \ldots, s$, we consider the rational function $g_i(x)/f(x) \in \mathbb{F}_q(x)$. Since $\deg(g_i) < m = \deg(f)$ by assumption, its formal Laurent series expansion has the form

$$\frac{g_i(x)}{f(x)} = \sum_{r=1}^{\infty} e_r^{(i)} x^{-r} \tag{4.54}$$

with coefficients $e_r^{(i)} \in \mathbb{F}_q$ for all $r \geq 1$. These coefficients serve as entries of the $m \times m$ generating matrix $C^{(i)} = (c_{j,k}^{(i)})_{1 \leq j \leq m,\, 0 \leq k \leq m-1}$ over $\mathbb{F}_q$. In detail, we set

$$c_{j,k}^{(i)} = e_{j+k}^{(i)} \in \mathbb{F}_q \qquad \text{for } 1 \leq i \leq s,\ 1 \leq j \leq m,\ 0 \leq k \leq m-1. \tag{4.55}$$

With these generating matrices $C^{(1)}, \ldots, C^{(s)}$, we apply the digital method and we obtain the point set $\mathcal{P}(\boldsymbol{g}, f)$ called a *polynomial lattice point set*. According to Corollary 4.4.21, $\mathcal{P}(\boldsymbol{g}, f)$ is a digital (t, m, s)-net over $\mathbb{F}_q$ with quality parameter

$$t = m - \varrho(C^{(1)}, \ldots, C^{(s)}).$$

We want to find out how the figure of merit $\varrho(C^{(1)}, \ldots, C^{(s)})$ depends on the inputs $\boldsymbol{g}$ and f of $\mathcal{P}(\boldsymbol{g}, f)$. According to Definitions 4.4.18 and 4.4.20, we have to study linear independence properties of the row vectors $\mathbf{c}_j^{(i)}$, $1 \leq j \leq m$, of the matrices $C^{(i)}$, $1 \leq i \leq s$. Note that

$$\mathbf{c}_j^{(i)} = (c_{j,0}^{(i)}, c_{j,1}^{(i)}, \ldots, c_{j,m-1}^{(i)}) \in \mathbb{F}_q^m \qquad \text{for } 1 \leq j \leq m,\ 1 \leq i \leq s, \tag{4.56}$$

where the coordinates $c_{j,k}^{(i)}$ are given by (4.55). For the sake of convenience, we introduce the "dot product" $\boldsymbol{h} \cdot \boldsymbol{g}$ for $\boldsymbol{h} = (h_1, \ldots, h_s) \in \mathbb{F}_q[x]^s$ by putting

$$\boldsymbol{h} \cdot \boldsymbol{g} = \sum_{i=1}^{s} h_i g_i \in \mathbb{F}_q[x].$$

Lemma 4.4.41 *The vectors $\mathbf{c}_j^{(i)}$ in (4.56) satisfy*

$$\sum_{i=1}^{s} \sum_{j=1}^{m} h_{i,j} \mathbf{c}_j^{(i)} = \mathbf{0} \in \mathbb{F}_q^m \tag{4.57}$$

with all $h_{i,j} \in \mathbb{F}_q$ if and only if f divides $\boldsymbol{h} \cdot \boldsymbol{g}$ in $\mathbb{F}_q[x]$, where $\boldsymbol{h} = (h_1, \ldots, h_s) \in \mathbb{F}_q[x]^s$ with

$$h_i(x) = \sum_{j=1}^{m} h_{i,j} x^{j-1} \in \mathbb{F}_q[x] \qquad \textit{for } 1 \leq i \leq s.$$

Proof. By comparing coordinates, we see that the linear dependence relation (4.57) is equivalent to

$$\sum_{i=1}^{s} \sum_{j=1}^{m} h_{i,j} e_{j+k}^{(i)} = 0 \qquad \text{for } 0 \leq k \leq m-1. \tag{4.58}$$

For each $i = 1, \ldots, s$, we obtain by (4.54) that

$$\begin{aligned} \frac{h_i(x) g_i(x)}{f(x)} &= \Big(\sum_{j=1}^{m} h_{i,j} x^{j-1} \Big) \Big(\sum_{r=1}^{\infty} e_r^{(i)} x^{-r} \Big) = \sum_{j=1}^{m} \sum_{r=1}^{\infty} h_{i,j} e_r^{(i)} x^{j-1-r} \\ &= \sum_{j=1}^{m} h_{i,j} \sum_{k=1-j}^{\infty} e_{j+k}^{(i)} x^{-k-1}. \end{aligned}$$

Thus for $k \geq 0$, the coefficient of x^{-k-1} in $h_i g_i / f$ is $\sum_{j=1}^{m} h_{i,j} e_{j+k}^{(i)}$. Therefore the condition (4.58) is equivalent to the property that for $0 \leq k \leq m-1$, the coefficient of x^{-k-1} in $\sum_{i=1}^{s} h_i g_i / f$ is 0. This means that

$$\frac{1}{f} \boldsymbol{h} \cdot \boldsymbol{g} = P + E,$$

where $P \in \mathbb{F}_q[x]$ and $E \in \mathbb{F}_q((x^{-1}))$ with $\nu(E) < -m$. The last identity is equivalent to

$$\boldsymbol{h} \cdot \boldsymbol{g} - Pf = Ef.$$

The left-hand side is a polynomial over $\mathbb{F}_q$, whereas on the right-hand side $\nu(Ef) = \nu(E) + \nu(f) < 0$ since $\nu(f) = \deg(f) = m$. This is possible if and only if $Ef = 0$, that is, if and only if f divides $\boldsymbol{h} \cdot \boldsymbol{g}$ in $\mathbb{F}_q[x]$. □

Remark 4.4.42 If integers $s \geq 2$ and $m \geq 1$ and an s-tuple $\boldsymbol{g} = (g_1, \ldots, g_s) \in \mathbb{F}_q[x]^s_{<m}$ are given, then there always exists a nonzero s-tuple $\boldsymbol{h} \in \mathbb{F}_q[x]^s_{<m}$ such that $\boldsymbol{h} \cdot \boldsymbol{g} = 0$. This is trivial if $g_i = 0$ for $1 \leq i \leq s$. If at least one g_i is nonzero, say without loss of generality $g_1 \neq 0$, then $\boldsymbol{h} = (g_2, -g_1, 0, \ldots, 0) \in \mathbb{F}_q[x]^s_{<m}$ is a suitable nonzero s-tuple $\boldsymbol{h}$. This simple argument shows that the minimum in the following theorem is extended over a nonempty set.

Theorem 4.4.43 *Let q be a prime power and let $s \geq 2$ and $m \geq 1$ be integers. Let $f \in \mathbb{F}_q[x]$ with* $\deg(f) = m$ *and let $\boldsymbol{g} \in \mathbb{F}_q[x]^s_{<m}$. Then the figure of merit of the generating matrices $C^{(1)}, \ldots, C^{(s)}$ of the polynomial lattice point set $\mathcal{P}(\boldsymbol{g}, f)$ is given by*

$$\varrho(C^{(1)}, \ldots, C^{(s)}) = \varrho(\boldsymbol{g}, f) := s - 1 + \min_{\boldsymbol{h}} \sum_{i=1}^{s} \deg(h_i),$$

where the minimum is extended over all nonzero s-tuples $\boldsymbol{h} = (h_1, \ldots, h_s) \in \mathbb{F}_q[x]^s_{<m}$ with f dividing $\boldsymbol{h} \cdot \boldsymbol{g}$ in $\mathbb{F}_q[x]$. Here we use the convention $\deg(0) = -1$.

Proof. By Definitions 4.4.18 and 4.4.20, there exist integers $d_1, \ldots, d_s$ with $0 \leq d_i \leq m$ for $1 \leq i \leq s$ and

$$\sum_{i=1}^{s} d_i = \varrho(C^{(1)}, \ldots, C^{(s)}) + 1$$

such that the system $\{\mathbf{c}_j^{(i)} \in \mathbb{F}_q^m : 1 \leq j \leq d_i,\ 1 \leq i \leq s\}$ is linearly dependent over $\mathbb{F}_q$. Hence there exist coefficients $h_{i,j} \in \mathbb{F}_q$, $1 \leq j \leq d_i$, $1 \leq i \leq s$, not all 0, such that

$$\sum_{i=1}^{s} \sum_{j=1}^{d_i} h_{i,j} \mathbf{c}_j^{(i)} = \mathbf{0} \in \mathbb{F}_q^m.$$

By putting $h_{i,j} = 0$ for $d_i < j \leq m$, $1 \leq i \leq s$, we obtain a linear dependence relation as in (4.57) in Lemma 4.4.41. This lemma implies that f divides $\boldsymbol{h} \cdot \boldsymbol{g}$ in $\mathbb{F}_q[x]$, with a nonzero s-tuple $\boldsymbol{h} = (h_1, \ldots, h_s) \in \mathbb{F}_q[x]^s$ as in the lemma. Note also that $\deg(h_i) \leq d_i - 1 < m$, and so the definition of $\varrho(\boldsymbol{g}, f)$ in the theorem shows that

$$\varrho(\boldsymbol{g}, f) \leq s - 1 + \sum_{i=1}^{s} \deg(h_i) \leq s - 1 + \sum_{i=1}^{s} (d_i - 1) = \varrho(C^{(1)}, \ldots, C^{(s)}).$$

In order to prove the converse inequality, we observe that by the definition of $\varrho(\boldsymbol{g}, f)$ in the theorem, there exists a nonzero s-tuple $\boldsymbol{h} = (h_1, \ldots, h_s) \in \mathbb{F}_q[x]^s_{<m}$

with f dividing $\boldsymbol{h} \cdot \boldsymbol{g}$ in $\mathbb{F}_q[x]$ and with

$$\varrho(\boldsymbol{g}, f) = s - 1 + \sum_{i-1}^{s} \deg(h_i).$$

Then Lemma 4.4.41 yields a linear dependence relation

$$\sum_{i=1}^{s} \sum_{j=1}^{d_i} h_{i,j} \mathbf{c}_j^{(i)} = \mathbf{0} \in \mathbb{F}_q^m,$$

where not all $h_{i,j} \in \mathbb{F}_q$ are 0 and where $d_i = \deg(h_i) + 1$ for $1 \leq i \leq s$ (here the convention $\deg(0) = -1$ is used). It follows now from the definition of $\varrho(C^{(1)}, \ldots, C^{(s)})$ that

$$\varrho(C^{(1)}, \ldots, C^{(s)}) \leq \sum_{i=1}^{s} d_i - 1 = \sum_{i=1}^{s} (\deg(h_i) + 1) - 1 = \varrho(\boldsymbol{g}, f),$$

and so we are done. □

Corollary 4.4.44 *Let q be a prime power and let $s \geq 2$ and $m \geq 1$ be integers. Let $f \in \mathbb{F}_q[x]$ with $\deg(f) = m$ and let $\boldsymbol{g} \in \mathbb{F}_q[x]_{<m}^s$. Then the polynomial lattice point set $\mathcal{P}(\boldsymbol{g}, f)$ is a digital (t, m, s)-net over $\mathbb{F}_q$ with*

$$t = m - \varrho(\boldsymbol{g}, f),$$

where $\varrho(\boldsymbol{g}, f)$ is as in Theorem 4.4.43.

Proof. This follows from Corollary 4.4.21 and Theorem 4.4.43. □

Example 4.4.45 Let q be an arbitrary prime power and let $f_0(x), f_1(x), \ldots$ be the sequence of Fibonacci polynomials over $\mathbb{F}_q$ defined recursively by $f_0(x) = 1$, $f_1(x) = x$, and $f_{k+2}(x) = x f_{k+1}(x) + f_k(x)$ for $k \geq 0$. It is obvious that $\deg(f_k) = k$ for all $k \geq 0$. Now for the dimension $s = 2$ and for an integer $m \geq 1$, we take $f = f_m$ and $\boldsymbol{g} = \boldsymbol{g}_m = (1, f_{m-1}) \in \mathbb{F}_q[x]_{<m}^2$ in the construction of polynomial lattice point sets. In this case by definition

$$\varrho(\boldsymbol{g}_m, f_m) = 1 + \min_{(h_1, h_2) \neq \mathbf{0}} (\deg(h_1) + \deg(h_2)),$$

where $(h_1, h_2) \in \mathbb{F}_q[x]_{<m}^2$ and f_m divides $h_1 + h_2 f_{m-1}$. We claim that $\varrho(\boldsymbol{g}_m, f_m) = m$ for all $m \geq 1$. Since we can take $(h_1, h_2) = (f_{m-1}, -1)$, it is clear that $\varrho(\boldsymbol{g}_m, f_m) \leq m$. Hence we want to prove that for every nonzero pair $(h_1, h_2) \in \mathbb{F}_q[x]_{<m}^2$ with f_m dividing $h_1 + h_2 f_{m-1}$, the inequality $\deg(h_1) + \deg(h_2) \geq m - 1$ is satisfied. This is trivial for $m = 1$. Now we proceed by induction and suppose that the assertion

is true for some $m \geq 1$. Then we consider a nonzero pair $(h_1, h_2) \in \mathbb{F}_q[x]^2$ with $\deg(h_1) \leq m$, $\deg(h_2) \leq m$, and f_{m+1} dividing $h_1 + h_2 f_m$. We have to verify that $\deg(h_1) + \deg(h_2) \geq m$. Since obviously $h_2 \neq 0$, we can assume that $\deg(h_1) \leq m-1$. Let us write $h_1 + h_2 f_m = P f_{m+1}$ for some $P \in \mathbb{F}_q[x]$. A comparison of degrees shows that $\deg(h_2) + m = \deg(P) + m + 1$, and so $\deg(P) = \deg(h_2) - 1$. Therefore $0 \leq \deg(P) \leq m-1$. Next we note that

$$h_1(x) + h_2(x) f_m(x) = P(x) f_{m+1}(x) = P(x) x f_m(x) + P(x) f_{m-1}(x),$$

and so f_m divides $h_1 - P f_{m-1}$. Hence by the induction hypothesis $\deg(h_1) + \deg(P) \geq m-1$, which implies that $\deg(h_1) + \deg(h_2) \geq m$ and completes the induction. It follows now from Corollary 4.4.44 that for all $m \geq 1$ the polynomial lattice point set $\mathcal{P}(\boldsymbol{g}_m, f_m)$ is a digital $(0, m, 2)$-net over $\mathbb{F}_q$. This example may be perceived as an analog for polynomial lattice point sets of the construction of two-dimensional good lattice points in Example 4.3.15.

For higher dimensions s, the situation for polynomial lattice point sets is similar to that for good lattice points and for hyperplane nets, meaning that there are theoretical existence results for good parameters, but no known explicit constructions of good polynomial lattice point sets. We establish such an existence result for good polynomial lattice point sets by using the same elimination method as in the proof of Theorem 4.4.38 for hyperplane nets. This similarity is not too surprising since it is known that polynomial lattice point sets belong to the family of hyperplane nets (see [39, Section 11.1]).

Theorem 4.4.46 *Let q be a prime power and let $s \geq 2$ and $m \geq 1$ be integers. Let the polynomial $f \in \mathbb{F}_q[x]$ be irreducible over $\mathbb{F}_q$ with* $\deg(f) = m$*. Then there exists an s-tuple $\boldsymbol{g} \in \mathbb{F}_q[x]^s_{<m}$ with*

$$\varrho(\boldsymbol{g}, f) \geq m - \lceil (s-1) \log_q (m+1) \rceil,$$

where $\varrho(\boldsymbol{g}, f)$ is as in Theorem 4.4.43 and where $\log_q$ *denotes the logarithm to the base q.*

Proof. We write

$$\varrho(\boldsymbol{g}, f) + 1 = \min_{\boldsymbol{h}} \sum_{i=1}^{s} (\deg(h_i) + 1),$$

where the minimum is extended over all nonzero s-tuples $\boldsymbol{h} = (h_1, \ldots, h_s) \in \mathbb{F}_q[x]^s_{<m}$ with f dividing $\boldsymbol{h} \cdot \boldsymbol{g}$ in $\mathbb{F}_q[x]$. We recall the convention $\deg(0) = -1$.

For every integer d with $0 \leq d \leq m$, the number of $h \in \mathbb{F}_q[x]$ with $\deg(h) + 1 = d$ is given by $\varepsilon_d q^d$, where $\varepsilon_0 = 1$ and $\varepsilon_d = (q-1)/q$ for $1 \leq d \leq m$. We can therefore use the same arguments as in the proof of Theorem 4.4.38 to show that for every integer k with $1 \leq k \leq m$, the number $B_q(k,m,s)$ of $\boldsymbol{h} = (h_1, \ldots, h_s) \in \mathbb{F}_q[x]^s$ with $\sum_{i=1}^s (\deg(h_i) + 1) = k$ satisfies

$$B_q(k,m,s) \leq (q-1)(m+1)^{s-1} q^{k-1}.$$

For a fixed $\boldsymbol{h} = (h_1, \ldots, h_s) \in \mathbb{F}_q[x]^s$ counted by $B_q(k,m,s)$, we determine the number of $\boldsymbol{g} = (g_1, \ldots, g_s) \in \mathbb{F}_q[x]^s_{<m}$ with f dividing $\boldsymbol{h} \cdot \boldsymbol{g}$ in $\mathbb{F}_q[x]$. Since f is irreducible over $\mathbb{F}_q$ by the hypothesis, the residue class field $\mathbb{F}_q[x]/(f(x))$ is a finite field of order q^m (compare with Remark 1.4.46). The condition that f divides $\boldsymbol{h} \cdot \boldsymbol{g}$ in $\mathbb{F}_q[x]$ can be expressed as the equation

$$\boldsymbol{h} \cdot \boldsymbol{g} = h_1 g_1 + \cdots + h_s g_s = 0$$

in $\mathbb{F}_q[x]/(f(x))$. From $1 \leq k \leq m$ we infer that $h_i \neq 0$ in $\mathbb{F}_q[x]/(f(x))$ for at least one i with $1 \leq i \leq s$, and since we are in a finite field of order q^m, it follows that the number of solutions $\boldsymbol{g} \in \mathbb{F}_q[x]^s_{<m}$ of $\boldsymbol{h} \cdot \boldsymbol{g} = 0$ is equal to $q^{m(s-1)}$. Therefore the number of $\boldsymbol{g} \in \mathbb{F}_q[x]^s_{<m}$ for which there exists an $\boldsymbol{h} \in \mathbb{F}_q[x]^s$ counted by $B_q(k,m,s)$ such that f divides $\boldsymbol{h} \cdot \boldsymbol{g}$ in $\mathbb{F}_q[x]$ is at most

$$B_q(k,m,s) q^{m(s-1)} \leq (q-1)(m+1)^{s-1} q^{m(s-1)+k-1}.$$

Now we put $K = m - \lceil (s-1) \log_q(m+1) \rceil$ as in the proof of Theorem 4.4.38 and we sum from $k = 1$ to $k = K$. Since there are q^{ms} candidate s-tuples $\boldsymbol{g} \in \mathbb{F}_q[x]^s_{<m}$, we conclude as in the proof of Theorem 4.4.38 that there is one such $\boldsymbol{g}_0 \in \mathbb{F}_q[x]^s_{<m}$ with $\varrho(\boldsymbol{g}_0, f) + 1 \geq K + 1$, that is, with $\varrho(\boldsymbol{g}_0, f) \geq K$. □

Corollary 4.4.47 *Let q be a prime power and let $s \geq 2$ and $m \geq 1$ be integers. Let the polynomial $f \in \mathbb{F}_q[x]$ be irreducible over $\mathbb{F}_q$ with* $\deg(f) = m$. *Then there exists an s-tuple $\boldsymbol{g} \in \mathbb{F}_q[x]^s_{<m}$ such that the polynomial lattice point set $\mathcal{P}(\boldsymbol{g}, f)$ is a digital (t,m,s)-net over $\mathbb{F}_q$ with*

$$t \leq \lceil (s-1) \log_q(m+1) \rceil.$$

Proof. This follows from Corollary 4.4.44 and Theorem 4.4.46. □

Corollary 4.4.47 has the same consequence for the star discrepancy of $\mathcal{P}(\boldsymbol{g}, f)$ as in the statement in the paragraph following the proof of Corollary 4.4.39. In the context of polynomial lattice point sets, the order of magnitude $N^{-1}(\log N)^{2s-2}$ in that discrepancy bound can be improved for $s \geq 3$ to $N^{-1}(\log N)^s$ by two different methods: a CBC algorithm for polynomial lattice point sets (see [39, Subsection 10.2.2]) and an analog for polynomial lattice point sets of the averaging technique in Theorem 4.3.7 (see [134, Theorem 4.43]).

We conclude the discussion of polynomial lattice point sets by presenting a description of these point sets which is completely analogous to that of the lattice point sets $\mathcal{P}(\mathbf{g},N)$ in Subsection 4.3.1. The essential tool is an analog of the map T_m in (4.48). For an integer $m \geq 1$, the map $U_m : \mathbb{F}_q((x^{-1})) \to [0,1)$ is given by

$$U_m(\sum_{r=w}^{\infty} e_r x^{-r}) = \sum_{r=\max(1,w)}^{m} \psi(e_r) q^{-r}$$

for every formal Laurent series $\sum_{r=w}^{\infty} e_r x^{-r} \in \mathbb{F}_q((x^{-1}))$, where $\psi : \mathbb{F}_q \to Z_q$ is the same bijection as in (4.48).

Theorem 4.4.48 *Let q be a prime power and let $s \geq 2$ and $m \geq 1$ be integers. Let $f \in \mathbb{F}_q[x]$ with $\deg(f) = m$ and let $\boldsymbol{g} = (g_1, \ldots, g_s) \in \mathbb{F}_q[x]^s_{<m}$. Then the polynomial lattice point set $\mathcal{P}(\boldsymbol{g}, f)$ consists of the points*

$$(U_m(vg_1/f), \ldots, U_m(vg_s/f)) \in [0,1)^s,$$

where v runs through the q^m polynomials in $\mathbb{F}_q[x]_{<m}$.

Proof. We fix an integer i with $1 \leq i \leq s$ and a polynomial $v(x) = \sum_{j=0}^{m-1} v_j x^j \in \mathbb{F}_q[x]_{<m}$ with $v_0, v_1, \ldots, v_{m-1} \in \mathbb{F}_q$. Using the expansion in (4.54), we obtain

$$\begin{aligned}\frac{v(x)g_i(x)}{f(x)} &= \Big(\sum_{j=0}^{m-1} v_j x^j\Big)\Big(\sum_{k=1}^{\infty} e_k^{(i)} x^{-k}\Big) = \sum_{j=0}^{m-1}\sum_{k=1}^{\infty} v_j e_k^{(i)} x^{j-k} \\ &= \sum_{r=2-m}^{\infty} \Big(\sum_{j=\max(1-r,0)}^{m-1} v_j e_{j+r}^{(i)}\Big) x^{-r}.\end{aligned}$$

The definition of the map U_m yields

$$U_m(vg_i/f) = \sum_{r=1}^{m} \psi\Big(\sum_{j=0}^{m-1} v_j e_{j+r}^{(i)}\Big) q^{-r}.$$

If we associate with the polynomial $v \in \mathbb{F}_q[x]_{<m}$ the column vector

$$\mathbf{v} = (v_0, v_1, \ldots, v_{m-1})^\top \in \mathbb{F}_q^m$$

and take into account (4.48) and the formula (4.55) for the entries of the generating matrix $C^{(i)}$ of $\mathcal{P}(\boldsymbol{g}, f)$, then we easily see that $U_m(vg_i/f) = T_m(C^{(i)}\mathbf{v})$. The proof is completed by referring to (4.49) and the fact that there is a one-to-one correspondence between the polynomials $v \in \mathbb{F}_q[x]_{<m}$ and the column vectors $\mathbf{v} \in \mathbb{F}_q^m$. □

The last construction of digital nets in this subsection is based on polynomial arithmetic over finite fields. These digital nets were introduced quite recently by

Hofer and Niederreiter [67] and they are called Vandermonde nets since their structure is reminiscent of that of Vandermonde matrices $(\alpha_i^{j-1})_{1\le i,j\le m}$ in linear algebra. The construction of Vandermonde nets works with the residue class ring $\mathbb{F}_q[x]/(f(x))$ for a polynomial $f(x)\in\mathbb{F}_q[x]$ of degree $m\ge 1$. We set up the map $\kappa_f:\mathbb{F}_q[x]\to\mathbb{F}_q^m$ as follows. Every $h\in\mathbb{F}_q[x]$ has a unique representative $\overline{h}\in\mathbb{F}_q[x]_{<m}$ in its residue class modulo f, namely the least residue $\overline{h}$ of h modulo f. Here $\mathbb{F}_q[x]_{<m}$ denotes as before the set of all polynomials over $\mathbb{F}_q$ of degree less than m. Now $\overline{h}(x)=\sum_{r=0}^{m-1}e_rx^r$ with $e_0,e_1,\dots,e_{m-1}\in\mathbb{F}_q$, and we put

$$\kappa_f(h)=(e_0,e_1,\dots,e_{m-1})\in\mathbb{F}_q^m.$$

It is obvious that κ_f is a linear transformation between vector spaces over $\mathbb{F}_q$.

The actual construction of Vandermonde nets proceeds by defining $m\times m$ generating matrices $C^{(1)},\dots,C^{(s)}$ over $\mathbb{F}_q$. We exclude the trivial one-dimensional case and assume that the dimension s satisfies $s\ge 2$. The basic constituents of a Vandermonde net are the same as for a polynomial lattice point set. We choose a polynomial $f\in\mathbb{F}_q[x]$ with $\deg(f)=m\ge 1$ and an s-tuple $\boldsymbol{g}=(g_1,\dots,g_s)\in\mathbb{F}_q[x]_{<m}^s$. The first generating matrix $C^{(1)}$ has the row vectors $\mathbf{c}_1^{(1)},\dots,\mathbf{c}_m^{(1)}$ with $\mathbf{c}_j^{(1)}=\kappa_f(g_1^{j-1})\in\mathbb{F}_q^m$ for $1\le j\le m$. For $i=2,\dots,s$, the jth row vector $\mathbf{c}_j^{(i)}$ of $C^{(i)}$ is given by $\mathbf{c}_j^{(i)}=\kappa_f(g_i^j)$ for $1\le j\le m$. The *Vandermonde net* $\mathcal{V}(\boldsymbol{g},f)$ is the digital net over $\mathbb{F}_q$ with generating matrices $C^{(1)},\dots,C^{(s)}$.

In order to determine the figure of merit $\varrho(C^{(1)},\dots,C^{(s)})$ (see Definition 4.4.20) of the generating matrices $C^{(1)},\dots,C^{(s)}$ of $\mathcal{V}(\boldsymbol{g},f)$, we introduce some more notation. We put

$$H_{q,m}=x\mathbb{F}_q[x]_{<m}=\{h\in\mathbb{F}_q[x]:\ \deg(h)\le m,\ h(0)=0\}.$$

Furthermore, we write $h\circ g$ for the composition of two polynomials $h,g\in\mathbb{F}_q[x]$, that is, $(h\circ g)(x)=h(g(x))$. Next, for $f\in\mathbb{F}_q[x]$ with $\deg(f)=m$ and $\boldsymbol{g}=(g_1,\dots,g_s)\in\mathbb{F}_q[x]_{<m}^s$ as above, we let $L(\boldsymbol{g},f)$ be the set of all s-tuples $\boldsymbol{h}=(h_1,\dots,h_s)\in\mathbb{F}_q[x]_{<m}\times H_{q,m}^{s-1}$ such that f divides $\sum_{i=1}^s(h_i\circ g_i)$ in $\mathbb{F}_q[x]$. Let us accentuate the interesting fact that there is some similarity with a condition in Theorem 4.4.43 for polynomial lattice point sets: there we had f dividing $\sum_{i=1}^s h_ig_i$ in $\mathbb{F}_q[x]$ and now we require that f divides $\sum_{i=1}^s(h_i\circ g_i)$ in $\mathbb{F}_q[x]$.

Remark 4.4.49 Under the usual conditions $s\ge 2$ and $m\ge 1$, the set $L(\boldsymbol{g},f)$ contains at least one nonzero s-tuple $\boldsymbol{h}$. In order to see this, we start from the obvious fact that the $m+1$ vectors $\kappa_f(1),\kappa_f(g_1),\kappa_f(g_1^2),\dots,\kappa_f(g_1^{m-1}),\kappa_f(g_2)$ in $\mathbb{F}_q^m$ must be linearly dependent over $\mathbb{F}_q$. Hence for some $b_0,b_1,\dots,b_m\in\mathbb{F}_q$, not all 0, we obtain

$$\sum_{j=0}^{m-1}b_j\kappa_f(g_1^j)+b_m\kappa_f(g_2)=\mathbf{0}\in\mathbb{F}_q^m.$$

Since κ_f is a linear transformation, this can also be written as

$$\kappa_f\Big(\sum_{j=0}^{m-1} b_j g_1^j + b_m g_2\Big) = \mathbf{0} \in \mathbb{F}_q^m.$$

The definition of κ_f implies that f divides $\sum_{j=0}^{m-1} b_j g_1^j + b_m g_2$ in $\mathbb{F}_q[x]$. Now we introduce the polynomials

$$h_1(x) = \sum_{j=0}^{m-1} b_j x^j, \quad h_2(x) = b_m x, \quad h_i(x) = 0 \quad \text{for } 3 \le i \le s.$$

Then $\boldsymbol{h} = (h_1, \dots, h_s) \in \mathbb{F}_q[x]_{<m} \times H_{q,m}^{s-1}$ is a nonzero s-tuple belonging to $L(\boldsymbol{g}, f)$.

In view of Remark 4.4.49, the set $L'(\boldsymbol{g}, f) := L(\boldsymbol{g}, f) \setminus \{\mathbf{0}\}$ is nonempty. We have to utilize two different degree functions on $\mathbb{F}_q[x]$, but they differ only in the way they are defined at the zero polynomial $0 \in \mathbb{F}_q[x]$. First, there is the standard degree function deg on $\mathbb{F}_q[x]$ with the convention $\deg(0) = -1$ in Theorem 4.4.43. Second, we use the degree function $\deg^*$ defined by $\deg^*(h) = \deg(h)$ for $h \in \mathbb{F}_q[x]$ with $h \neq 0$ and $\deg^*(0) = 0$.

Theorem 4.4.50 *Let q be a prime power and let $s \ge 2$ and $m \ge 1$ be integers. Let $f \in \mathbb{F}_q[x]$ with $\deg(f) = m$ and let $\boldsymbol{g} \in \mathbb{F}_q[x]_{<m}^s$. Then the figure of merit of the generating matrices $C^{(1)}, \dots, C^{(s)}$ of the Vandermonde net $\mathcal{V}(\boldsymbol{g}, f)$ is given by*

$$\varrho(C^{(1)}, \dots, C^{(s)}) = \mu(\boldsymbol{g}, f) := \min_{(h_1,\dots,h_s) \in L'(\boldsymbol{g},f)} \Big(\deg(h_1) + \sum_{i=2}^{s} \deg^*(h_i)\Big).$$

Proof. Let $d_1, \dots, d_s$ be integers with $0 \le d_i \le m$ for $1 \le i \le s$ and $\sum_{i=1}^{s} d_i \ge 1$ such that there exists a linear dependence relation

$$\sum_{i=1}^{s} \sum_{j=1}^{d_i} b_{i,j} \mathbf{c}_j^{(i)} = \mathbf{0} \in \mathbb{F}_q^m, \tag{4.59}$$

where all $b_{i,j} \in \mathbb{F}_q$ and not all of them are 0. Here we can assume that $b_{i,d_i} \neq 0$ if $d_i \ge 1$. By the definition of the row vectors $\mathbf{c}_j^{(i)}$ of the generating matrices, (4.59) is equivalent to

$$\sum_{j=1}^{d_1} b_{1,j} \kappa_f(g_1^{j-1}) + \sum_{i=2}^{s} \sum_{j=1}^{d_i} b_{i,j} \kappa_f(g_i^j) = \mathbf{0} \in \mathbb{F}_q^m.$$

Because of the linearity of κ_f, this is in turn equivalent to

$$\kappa_f\Big(\sum_{j=1}^{d_1} b_{1,j} g_1^{j-1} + \sum_{i=2}^{s} \sum_{j=1}^{d_i} b_{i,j} g_i^j\Big) = \mathbf{0} \in \mathbb{F}_q^m.$$

This means that f divides $\sum_{i=1}^{s}(h_i \circ g_i)$ in $\mathbb{F}_q[x]$, where

$$h_1(x) = \sum_{j=1}^{d_1} b_{1,j} x^{j-1} \in \mathbb{F}_q[x]_{<m}, \quad h_i(x) = \sum_{j=1}^{d_i} b_{i,j} x^j \in H_{q,m} \quad \text{for } 2 \leq i \leq s.$$

Therefore (4.59) is equivalent to $\boldsymbol{h} = (h_1, \ldots, h_s)$ belonging to $L'(\boldsymbol{g}, f)$. Furthermore, by the definitions of the degree functions deg and deg*, it is evident that $\deg(h_1) = d_1 - 1$ and $\deg^*(h_i) = d_i$ for $2 \leq i \leq s$, and so

$$\deg(h_1) + \sum_{i=2}^{s} \deg^*(h_i) = \sum_{i=1}^{s} d_i - 1. \tag{4.60}$$

By the definition of $\varrho(C^{(1)}, \ldots, C^{(s)})$, there exist $d_1, \ldots, d_s$ as above with $\sum_{i=1}^{s} d_i = \varrho(C^{(1)}, \ldots, C^{(s)}) + 1$. Consequently, there is an $\boldsymbol{h} \in L'(\boldsymbol{g}, f)$ with the left-hand side of (4.60) equal to $\varrho(C^{(1)}, \ldots, C^{(s)})$. Hence the definition of $\mu(\boldsymbol{g}, f)$ in the theorem implies that $\mu(\boldsymbol{g}, f) \leq \varrho(C^{(1)}, \ldots, C^{(s)})$. Conversely, there exists an $\boldsymbol{h} \in L'(\boldsymbol{g}, f)$ with the left-hand side of (4.60) equal to $\mu(\boldsymbol{g}, f)$. This yields a linear dependence relation (4.59) with $\sum_{i=1}^{s} d_i = \mu(\boldsymbol{g}, f) + 1$, and so $\varrho(C^{(1)}, \ldots, C^{(s)}) \leq \mu(\boldsymbol{g}, f)$. □

Corollary 4.4.51 *Let q be a prime power and let $s \geq 2$ and $m \geq 1$ be integers. Let $f \in \mathbb{F}_q[x]$ with $\deg(f) = m$ and let $\boldsymbol{g} \in \mathbb{F}_q[x]_{<m}^s$. Then the Vandermonde net $\mathcal{V}(\boldsymbol{g}, f)$ is a digital (t, m, s)-net over $\mathbb{F}_q$ with*

$$t = m - \mu(\boldsymbol{g}, f),$$

where $\mu(\boldsymbol{g}, f)$ is as in Theorem 4.4.50.

Proof. This follows from Corollary 4.4.21 and Theorem 4.4.50. □

Remark 4.4.52 You may wonder why there is a certain asymmetry in the definition of the row vectors $\mathbf{c}_j^{(i)}$ of the generating matrices of $\mathcal{V}(\boldsymbol{g}, f)$. Remember that we defined $\mathbf{c}_j^{(1)} = \kappa_f(g_1^{j-1})$ for $1 \leq j \leq m$, but $\mathbf{c}_j^{(i)} = \kappa_f(g_i^j)$ for $2 \leq i \leq s$ and $1 \leq j \leq m$. A symmetric definition, which would also produce a perfect Vandermonde structure, would be $\mathbf{c}_j^{(i)} = \kappa_f(g_i^{j-1})$ for $1 \leq i \leq s$ and $1 \leq j \leq m$. But with this definition we would obtain

$$\mathbf{c}_1^{(i)} = \kappa_f(1) = (1, 0, \ldots, 0) \in \mathbb{F}_q^m \qquad \text{for } 1 \leq i \leq s,$$

and so in particular $\mathbf{c}_1^{(1)}$ and $\mathbf{c}_1^{(2)}$ would be linearly dependent over $\mathbb{F}_q$. Therefore $\varrho(C^{(1)}, \ldots, C^{(s)}) = 1$, and we would get an uninteresting digital $(m-1, m, s)$-net over $\mathbb{F}_q$.

There are several parallels between Vandermonde nets and polynomial lattice point sets. For instance, there is an existence result for large values of the figure of merit $\mu(\boldsymbol{g}, f)$ comparable to Theorem 4.4.46, there is an existence theorem for Vandermonde nets with small star discrepancy by an averaging technique, and there is a CBC algorithm for computing parameters of good Vandermonde nets. All this can be found in the paper [67].

We have one more card up our sleeve, namely an explicit construction of Vandermonde nets with optimal quality parameter $t = 0$ that goes beyond the two-dimensional case. For polynomial lattice point sets, an explicit construction yielding $t = 0$ for arbitrary m and q is known only for the dimension $s = 2$ (see Example 4.4.45). This explicit construction of Vandermonde nets is also due to Hofer and Niederreiter [67] and proceeds as follows. Let q be an arbitrary prime power, let s be a dimension with $2 \leq s \leq q + 1$, and let $m \geq 2$ be an integer. As usual, we choose a polynomial $f \in \mathbb{F}_q[x]$ with $\deg(f) = m$. For simplicity we assume that f is irreducible over $\mathbb{F}_q$, but the astute reader will observe that more general choices of f are possible as well. Next we select $s - 1$ distinct elements $c_2, \dots, c_s$ of $\mathbb{F}_q$. This is feasible since $s - 1 \leq q$. Finally, we put $g_1(x) = x \in \mathbb{F}_q[x]_{<m}$ and, for each $i = 2, \dots, s$, let $g_i(x) \in \mathbb{F}_q[x]_{<m}$ be the polynomial that is uniquely determined by the congruence

$$g_i(x)(x - c_i) \equiv 1 \pmod{f(x)}.$$

Note that $g_i(x)$ exists since $\mathbb{F}_q[x]/(f(x))$ is a field. With these polynomials $g_1, \dots, g_s$, we set up the s-tuple $\boldsymbol{g} = (g_1, \dots, g_s) \in \mathbb{F}_q[x]^s_{<m}$.

Theorem 4.4.53 *Let q be a prime power and let s and m be integers with $2 \leq s \leq q + 1$ and $m \geq 2$. Let $f \in \mathbb{F}_q[x]$ be irreducible over $\mathbb{F}_q$ with $\deg(f) = m$ and let $\boldsymbol{g} \in \mathbb{F}_q[x]^s_{<m}$ be the s-tuple of polynomials constructed above. Then the Vandermonde net $\mathcal{V}(\boldsymbol{g}, f)$ is a digital $(0, m, s)$-net over $\mathbb{F}_q$.*

Proof. According to Corollary 4.4.51, we have to show that $\mu(\boldsymbol{g}, f) = m$. We prove this by contradiction and suppose that $\mu(\boldsymbol{g}, f) \leq m - 1$. Then by the definition of $\mu(\boldsymbol{g}, f)$ in Theorem 4.4.50, there exists an s-tuple $\boldsymbol{h} = (h_1, \dots, h_s) \in L'(\boldsymbol{g}, f)$ with $\sum_{i=1}^{s} d_i \leq m - 1$, where $d_1 = \deg(h_1)$ and $d_i = \deg^*(h_i)$ for $2 \leq i \leq s$. Put $h_i(x) = \sum_{j=1}^{d_i} h_{i,j} x^j$ for $2 \leq i \leq s$ with all $h_{i,j} \in \mathbb{F}_q$. Then $f(x)$ divides

$$\sum_{i=1}^{s} (h_i \circ g_i)(x) = h_1(x) + \sum_{i=2}^{s} \sum_{j=1}^{d_i} h_{i,j} g_i(x)^j$$

in $\mathbb{F}_q[x]$. Multiplying by $\prod_{k=2}^{s} (x - c_k)^{d_k}$, we deduce that $f(x)$ divides

$$h_1(x) \prod_{k=2}^{s} (x - c_k)^{d_k} + \sum_{i=2}^{s} \Big(\sum_{j=1}^{d_i} h_{i,j} g_i(x)^j \Big) (x - c_i)^{d_i} \prod_{\substack{k=2 \\ k \neq i}}^{s} (x - c_k)^{d_k}$$

in $\mathbb{F}_q[x]$. Now $g_i(x)^j(x-c_i)^j \equiv 1 \pmod{f(x)}$ for $2 \leq i \leq s$ and $j \geq 1$, and so by working with congruences modulo $f(x)$ we see that $f(x)$ divides

$$M(x) := h_1(x)\prod_{k=2}^{s}(x-c_k)^{d_k} + \sum_{i=2}^{s}\Big(\sum_{j=1}^{d_i} h_{i,j}(x-c_i)^{d_i-j}\Big)\prod_{\substack{k=2\\k\neq i}}^{s}(x-c_k)^{d_k}$$

in $\mathbb{F}_q[x]$. Now we consider $\deg(M)$. The first term of $M(x)$ has degree $\leq \sum_{k=1}^{s} d_k \leq m-1$. In the sum $\sum_{i=2}^{s}$ in the expression for $M(x)$, a term appears only if $d_i \geq 1$ and such a term has degree $\leq \sum_{k=2}^{s} d_k - 1 \leq \sum_{k=1}^{s} d_k \leq m-1$ since $d_1 = \deg(h_1) \geq -1$. Altogether we obtain $\deg(M) \leq m-1 < \deg(f)$. Since $f(x)$ divides $M(x)$ in $\mathbb{F}_q[x]$, it follows that $M(x)$ is the zero polynomial in $\mathbb{F}_q[x]$. If we assume that $d_r \geq 1$ for some $r \in \{2,\ldots,s\}$, then substituting $x = c_r$ in $M(x)$ we get

$$0 = M(c_r) = \sum_{j=1}^{d_r} h_{r,j}(c_r-c_r)^{d_r-j}\prod_{\substack{k=2\\k\neq r}}^{s}(c_r-c_k)^{d_k} = h_{r,d_r}\prod_{\substack{k=2\\k\neq r}}^{s}(c_r-c_k)^{d_k}.$$

Since the last product is nonzero, we deduce that $h_{r,d_r} = 0$. But this is a contradiction to $\deg^*(h_r) = d_r$. Thus we have shown that $d_i = 0$ for $2 \leq i \leq s$, and so $h_i = 0 \in \mathbb{F}_q[x]$ for $2 \leq i \leq s$. Since $M = 0 \in \mathbb{F}_q[x]$, it follows that also $h_1 = 0 \in \mathbb{F}_q[x]$. This is the final contradiction, since $\boldsymbol{h} \in L'(\boldsymbol{g}, f)$ means in particular that $\boldsymbol{h}$ is a nonzero s-tuple. □

Remark 4.4.54 In principle, the construction in Theorem 4.4.53 works also in the case $s = 1$. The vectors $\kappa_f(g_1^{j-1}) = \kappa_f(x^{j-1}) \in \mathbb{F}_q^m$ for $1 \leq j \leq m$ are then just the row vectors of the $m \times m$ identity matrix over $\mathbb{F}_q$, and so the construction coincides with that in Example 4.4.16. The remarkable fact about Theorem 4.4.53 is that the condition $s \leq q+1$ is best possible. This follows from Theorem 4.4.11 which implies that if $m \geq 2$, then a $(0,m,s)$-net in base q can exist only if $s \leq q+1$. Thus, for a prime power q and an integer $m \geq 2$, we can say that a $(0,m,s)$-net in base q exists if and only if $s \leq q+1$, and if it exists, then we can even construct a digital $(0,m,s)$-net over $\mathbb{F}_q$.

4.4.4 (t,s)-sequences

We have seen that the theory of (t,m,s)-nets provides a systematic way of obtaining point sets with small discrepancy. An analogous approach to the construction of low-discrepancy sequences is afforded by the theory of (t,s)-sequences. In a nutshell: what (t,m,s)-nets are for point sets, (t,s)-sequences are for (infinite) sequences. Note that in this chapter we have not yet advanced very far with the construction of low-discrepancy sequences in the sense of the discrepancy bound (4.22). Up to now we encountered only one construction of low-discrepancy sequences that works in any dimension, namely that of Halton sequences in Subsection 4.2.2. Many more examples of low-discrepancy sequences are furnished by the theory of (t, s)-sequences.

Informally, a (t, s)-sequence in base b is a sequence of points in $[0, 1)^s$ such that certain blocks of terms form (t, m, s)-nets in base b with t independent of m. Since a (t, m, s)-net in base b consists of b^m points, the lengths of the blocks taken from the sequence must be powers of b. The formal definition is as follows. We again use the abbreviation $(\mathbf{x}_n)_{n=0}^{\infty}$ for the sequence of points $\mathbf{x}_0, \mathbf{x}_1, \ldots$.

Definition 4.4.55 Let $b \geq 2$, $s \geq 1$, and $t \geq 0$ be integers. A (t, s)-*sequence in base* b is a sequence $(\mathbf{x}_n)_{n=0}^{\infty}$ of points in $[0, 1)^s$ with the property that, for all integers $k \geq 0$ and $m > t$, the point set consisting of the $\mathbf{x}_n$ with $kb^m \leq n < (k + 1)b^m$ is a (t, m, s)-net in base b.

Example 4.4.56 For an arbitrary base $b \geq 2$ and the dimension $s = 1$, let us consider the van der Corput sequence $(x_n)_{n=0}^{\infty}$ in base b given by $x_n = \phi_b(n)$ for all $n \geq 0$ (see Remark 4.2.9). We claim that this sequence is a $(0, 1)$-sequence in base b. We prove this by taking integers $k \geq 0$ and $m \geq 1$ and looking at the b^m points x_n with $kb^m \leq n < (k+1)b^m$. We have to show that these b^m points form a $(0, m, 1)$-net in base b. We proceed by the definition of such a net in Definition 4.4.4 and consider any one-dimensional elementary interval $J = [ab^{-m}, (a + 1)b^{-m})$ in base b with $a \in \mathbb{Z}$ and $0 \leq a < b^m$. Note that $\phi_b(n) \in [ab^{-m}, (a + 1)b^{-m})$ means that the first m b-adic digits of $\phi_b(n)$ are prescribed, or equivalently that in the digit expansion $n = \sum_{j=0}^{\infty} z_j(n)b^j$ of n in (4.30) the digits $z_0(n), z_1(n), \ldots, z_{m-1}(n)$ are prescribed. But in the range $kb^m \leq n < (k + 1)b^m$ there is exactly one value of n with these prescribed digits, and so J contains exactly one point x_n with $kb^m \leq n < (k+1)b^m$. Thus, we have verified that the van der Corput sequence in base b is a $(0, 1)$-sequence in base b. Historically, the van der Corput sequence in base b served as the model for the definition of a (t, s)-sequence in base b.

As for (t, m, s)-nets in base b, the parameter t of a (t, s)-sequence in base b is called the *quality parameter*, and we want this parameter to be as small as possible. We have trivial consequences of two results in Subsection 4.4.1. First, Proposition 4.4.8 implies that if $b \geq 2$, $s \geq 1$, and $t \geq 0$ are integers and if $\mathcal{S}$ is a (t, s)-sequence in base b, then $\mathcal{S}$ is also a (v, s)-sequence in base b for every integer $v \geq t$. Second, Proposition 4.4.9 shows that if $b \geq 2$, $s \geq 2$, and $t \geq 0$ are integers and if a (t, s)-sequence in base b is given, then by projection we get a (t, r)-sequence in base b for every dimension r with $1 \leq r < s$.

It is inherent in Definition 4.4.55 that once we get hold of a (t, s)-sequence in base b, then we automatically obtain (t, m, s)-nets in base b for infinitely many values of m. By a simple trick that we already used in Lemma 4.1.38 and (4.32), we can actually construct infinitely many $(s + 1)$-dimensional nets.

Proposition 4.4.57 *Let $b \geq 2$, $s \geq 1$, and $t \geq 0$ be integers. If $(\mathbf{x}_n)_{n=0}^{\infty}$ is a (t, s)-sequence in base b, then for every integer $m \geq t$ the points*

$$\mathbf{y}_n = (nb^{-m}, \mathbf{x}_n) \in [0,1)^{s+1} \qquad \text{for } n = 0, 1, \ldots, b^m - 1$$

form a $(t, m, s+1)$-net in base b.

Proof. Let $\mathcal{P}$ be the point set consisting of $\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_{b^m-1}$. Let

$$J = \prod_{i=1}^{s+1} [a_i b^{-d_i}, (a_i + 1) b^{-d_i}) \subseteq [0,1)^{s+1}$$

with $a_i, d_i \in \mathbb{Z}$, $d_i \geq 0$, and $0 \leq a_i < b^{d_i}$ for $1 \leq i \leq s+1$ be an $(s+1)$-dimensional elementary interval in base b with $\lambda_{s+1}(J) = b^{t-m}$, that is, with $\sum_{i=1}^{s+1} d_i = m - t$. Now $\mathbf{y}_n \in J$ if and only if $a_1 b^{m-d_1} \leq n < (a_1 + 1) b^{m-d_1}$ and

$$\mathbf{x}_n \in J' := \prod_{i=2}^{s+1} [a_i b^{-d_i}, (a_i + 1) b^{-d_i}).$$

Since $m - d_1 = t + \sum_{i=2}^{s+1} d_i \geq t$ and $(\mathbf{x}_n)_{n=0}^{\infty}$ is a (t, s)-sequence in base b, the point set $\mathcal{P}'$ consisting of the $\mathbf{x}_n$ with $a_1 b^{m-d_1} \leq n < (a_1 + 1) b^{m-d_1}$ is a $(t, m - d_1, s)$-net in base b (note that this holds trivially if $m - d_1 = t$). Now J' is an s-dimensional elementary interval in base b with $\lambda_s(J') = b^{t-(m-d_1)}$, and so $A(J'; \mathcal{P}') = b^t$ by the definition of a $(t, m - d_1, s)$-net in base b. Therefore $A(J; \mathcal{P}) = b^t$ as desired. □

Theorem 4.4.58 *Let $b \geq 2$ be an integer. Then a $(0, s)$-sequence in base b can exist only if $s \leq b$.*

Proof. Let $s \geq 1$ be a dimension for which there exists a $(0, s)$-sequence in base b. Then Proposition 4.4.57 yields a $(0, 2, s+1)$-net in base b. It follows now from Theorem 4.4.11 that $s + 1 \leq b + 1$, and so $s \leq b$. □

A discrepancy bound for (t, m, s)-nets in base b was formulated in Theorem 4.4.14. By using this bound and Definition 4.4.55, one can derive a discrepancy bound for (t, s)-sequences in base b. The following bound, which we state without proof, is obtained by combining results from [89] and [50].

Theorem 4.4.59 *Let $b \geq 2$, $s \geq 1$, and $t \geq 0$ be integers. Then the star discrepancy $D_N^*(\mathcal{S})$ of a (t, s)-sequence $\mathcal{S}$ in base b satisfies*

$$N D_N^*(\mathcal{S}) \leq \frac{\lfloor b^2/2 \rfloor}{b^2 - 1} \cdot \frac{b^t}{s!} \Big(\frac{b-1}{2 \log b} \Big)^s (\log N)^s + C(b, s) b^t (\log N)^{s-1}$$

for all $N \geq 2$, where the constant $C(b, s) > 0$ depends only on b and s.

It is a striking consequence of the discrepancy bound in Theorem 4.4.59 that *any* (t, s)-sequence in base b is a low-discrepancy sequence in the technical sense of (4.22), and so is in particular uniformly distributed in $[0, 1)^s$ by Theorem 4.1.36. Theorem 4.4.59 indicates again that small values of the quality parameter t are preferable in a (t, s)-sequence in base b. If the base $b \geq 2$ is fixed, then Theorem 4.4.58 shows, regrettably, that the optimal value $t = 0$ can be achieved only for finitely many dimensions s.

The only example of a (t, s)-sequence in base b that we have encountered so far is the simple $(0, 1)$-sequence in base b described in Example 4.4.56. In general, the construction of (t, s)-sequences in base b for dimensions $s \geq 2$ is a challenging task since the requirements in Definition 4.4.55 are quite severe. What we need, first of all, is a systematic way of approaching the problem of constructing (t, s)-sequences, and we take our cue from the digital method for the construction of nets presented in Subsection 4.4.2. There the basic ingredients were generating matrices over a finite field.

The digital method for the construction of (t, s)-sequences works again with generating matrices, but now the generating matrices are of infinite size, in line with the aim that we want to construct an infinite sequence rather than a finite point set. What complicates matters further is the fact that we need to exercise more care about the order in which the points of the sequence are listed. For a finite point set and its (star) discrepancy, the order of the points is irrelevant. For an infinite sequence, not only its (star) discrepancy but even its distribution properties depend a lot on the way the points of the sequence are listed, as the following easy example demonstrates.

Example 4.4.60 Let $\mathcal{S} = (x_n)_{n=1}^{\infty}$ be any uniformly distributed sequence of distinct points in $[0, 1)$, such as a Kronecker sequence. Then Theorem 4.1.6 implies that there are infinitely many x_n in the interval $[0, \frac{1}{2})$ and infinitely many x_n in the interval $[\frac{1}{2}, 1)$. Let $n_1 < n_2 < \cdots$ be all those subscripts n for which $x_n \in [0, \frac{1}{2})$ and let $m_1 < m_2 < \cdots$ be all those subscripts n for which $x_n \in [\frac{1}{2}, 1)$. Now we rearrange the sequence $\mathcal{S}$ into a sequence $\mathcal{S}' = (y_n)_{n=1}^{\infty}$ as follows. We put $y_1 = x_{n_1}$, $y_2 = x_{n_2}$, $y_3 = x_{m_1}$, $y_4 = x_{n_3}$, $y_5 = x_{n_4}$, $y_6 = x_{m_2}$, and so on in an obvious fashion; that is, we always pick the two still unused $x_n \in [0, \frac{1}{2})$ with the least subscripts and then the one still unused $x_n \in [\frac{1}{2}, 1)$ with the least subscript. Then it is obvious that for every integer $N \geq 1$ and for $J = [0, \frac{1}{2})$ we get $\sum_{n=1}^{3N} c_J(y_n) = 2N$, and so

$$\lim_{N \to \infty} \frac{1}{3N} \sum_{n=1}^{3N} c_J(y_n) = \frac{2}{3} \neq \lambda(J).$$

Hence by Theorem 4.1.6, the sequence $\mathcal{S}'$ is not uniformly distributed in $[0, 1)$. Note that $\mathcal{S}$ and $\mathcal{S}'$ coincide as sets, but they differ as sequences since the same points are listed in a different order. The lesson of this example is that a changed order can completely change the distribution behavior.

After these preparations, we get down to business and describe the *digital method* for the construction of (t, s)-sequences in detail. Let q be an arbitrary prime power and let $\mathbb{F}_q$ be the finite field of order q. Let $\mathbb{F}_q^\omega$ be the set of all sequences of elements of $\mathbb{F}_q$ with only finitely many nonzero terms. We think of these sequences also as column vectors of infinite length. We set up the map $T_\infty : \mathbb{F}_q^\omega \to [0, 1)$, which is an analog of the map T_m in (4.48), by putting

$$T_\infty(\mathbf{h}) = \sum_{j=1}^{\infty} \psi(h_j) q^{-j} \tag{4.61}$$

for every column vector $\mathbf{h} = (h_1, h_2, \ldots)^\top \in \mathbb{F}_q^\omega$, where $\psi : \mathbb{F}_q \to Z_q$ is a fixed bijection with $\psi(0) = 0$. Note that $T_\infty(\mathbf{h})$ is always a number in $[0, 1)$ with a finite digit expansion in base q.

For a given dimension $s \geq 1$, we choose $\infty \times \infty$ matrices $C^{(1)}, \ldots, C^{(s)}$ over $\mathbb{F}_q$, where by an $\infty \times \infty$ matrix we mean a matrix with denumerably many rows and columns. Each matrix $C^{(i)}$, $i = 1. \ldots, s$, must have the property that each column of $C^{(i)}$ contains only finitely many nonzero entries. Remember that we want to construct a sequence $\mathbf{x}_0, \mathbf{x}_1, \ldots$ of points in $[0, 1)^s$. In order to define $\mathbf{x}_n$, we take an integer $n \geq 0$ and let

$$n = \sum_{r=1}^{\infty} z_r(n) q^{r-1} \tag{4.62}$$

be its unique digit expansion in base q, where $z_r(n) \in Z_q$ for all $r \geq 1$ and $z_r(n) = 0$ for all sufficiently large r. Next we fix a bijection $\eta : Z_q \to \mathbb{F}_q$ with $\eta(0) = 0$. Then we associate with the integer n the column vector

$$\mathbf{n} = (\eta(z_1(n)), \eta(z_2(n)), \ldots)^\top$$

of infinite length. Because of the conditions on the $z_r(n)$ and on η, we can guarantee that $\mathbf{n} \in \mathbb{F}_q^\omega$.

For a fixed i with $1 \leq i \leq s$, we form now the matrix-vector product $C^{(i)}\mathbf{n}$. This is again a column vector of infinite length and its jth entry is the "dot product" of the jth row of $C^{(i)}$ with $\mathbf{n}$. This "dot product", defined in analogy with Definition 3.2.17, is formally an infinite sum. This may cause a problem since there is no concept of convergence in $\mathbb{F}_q$, but the fact that $\eta(z_r(n)) = 0$ for all sufficiently large r implies that the "dot product" is just a finite sum and therefore meaningful. We actually want to make sure that $C^{(i)}\mathbf{n}$ belongs to $\mathbb{F}_q^\omega$; recall that this means that this column vector has only finitely many nonzero coordinates. Let $r_0 \in \mathbb{N}$ be such that $\eta(z_r(n)) = 0$ for all $r > r_0$. For all $j \geq 1$, the jth coordinate of $C^{(i)}\mathbf{n}$ is the "dot product" of the jth row $(c_{j1}^{(i)}, c_{j2}^{(i)}, \ldots)$ of $C^{(i)}$ with $\mathbf{n}$, and so it is given by

$$\sum_{r=1}^{r_0} c_{jr}^{(i)} \eta(z_r(n)) \in \mathbb{F}_q. \tag{4.63}$$

Now we consider the first r_0 columns of $C^{(i)}$. By assumption, each of these columns contains only finitely many nonzero entries, and so there exists a $j_0 \in \mathbb{N}$ such that $c_{jr}^{(i)} = 0$ for all $j \geq j_0$ and $1 \leq r \leq r_0$. Consequently, the element in (4.63) is equal to 0 for $j \geq j_0$, and so we get indeed $C^{(i)}\mathbf{n} \in \mathbb{F}_q^{\omega}$. Therefore it makes sense to define

$$\mathbf{x}_n = (T_\infty(C^{(1)}\mathbf{n}), \ldots, T_\infty(C^{(s)}\mathbf{n})) \in [0,1)^s \qquad \text{for } n = 0, 1, \ldots, \tag{4.64}$$

where T_∞ is the map in (4.61).

Definition 4.4.61 The sequence $\mathcal{S} = (\mathbf{x}_n)_{n=0}^{\infty}$ defined in (4.64) is called a *digital sequence over* $\mathbb{F}_q$ and the matrices $C^{(1)}, \ldots, C^{(s)}$ are the *generating matrices* of $\mathcal{S}$. If $\mathcal{S}$ is a (t, s)-sequence in base q for some integer $t \geq 0$, then $\mathcal{S}$ is called a *digital* (t, s)*-sequence over* $\mathbb{F}_q$.

Example 4.4.62 Let $s = 1$, let q be an arbitrary prime power, and let the bijections $\psi : \mathbb{F}_q \to Z_q$ and $\eta : Z_q \to \mathbb{F}_q$ be inverse maps of each other. We choose the generating matrix $C^{(1)}$ to be the $\infty \times \infty$ identity matrix over $\mathbb{F}_q$ which is defined in the obvious fashion. Then it is easily seen that the corresponding digital sequence over $\mathbb{F}_q$ is the van der Corput sequence in base q. We learned in Example 4.4.56 that this sequence is a $(0, 1)$-sequence in base q, and so the van der Corput sequence in base q is a digital $(0, 1)$-sequence over $\mathbb{F}_q$.

Every s-dimensional digital net over $\mathbb{F}_q$ with $m \times m$ generating matrices is a digital (t, m, s)-net over $\mathbb{F}_q$ for some value of t, in the worst case for $t = m$. The analogous statement for digital sequences over $\mathbb{F}_q$ is not correct. There are bad choices of the $\infty \times \infty$ generating matrices that do not produce (t, s)-sequences in base q, no matter how large we make t; for instance, let $s \geq 2$ and let the s generating matrices all be equal. The following theorem provides insight into the condition that the $\infty \times \infty$ generating matrices $C^{(1)}, \ldots, C^{(s)}$ have to satisfy in order to obtain a digital (t, s)-sequence over $\mathbb{F}_q$. For $i = 1, \ldots, s$ and every integer $m \geq 1$, we write $C_m^{(i)}$ for the upper left $m \times m$ submatrix of $C^{(i)}$. Furthermore, we use the figure of merit introduced in Definition 4.4.20.

Theorem 4.4.63 *Let q be a prime power, let $s \geq 1$ be an integer, and let $C^{(1)}, \ldots, C^{(s)}$ be the $\infty \times \infty$ generating matrices over $\mathbb{F}_q$ of a digital sequence $\mathcal{S}$ over $\mathbb{F}_q$. If*

$$t := \sup_{m \in \mathbb{N}} \left(m - \varrho(C_m^{(1)}, \ldots, C_m^{(s)})\right)$$

is finite, then $\mathcal{S}$ is a digital (t, s)-sequence over $\mathbb{F}_q$.

Proof. We have to verify the various net properties in Definition 4.4.55, and so we fix integers $k \geq 0$ and $m > t$ and we consider the points $\mathbf{x}_n$ in (4.64) with $kq^m \leq n < (k+1)q^m$. In this range, the q-adic digits $z_r(n)$ of n in (4.62) are prescribed for $r > m$, whereas the $z_r(n)$ with $1 \leq r \leq m$ can range freely over Z_q. In order to prove the desired net property, we take an elementary interval

$$J = \prod_{i=1}^{s} [a_i q^{-d_i}, (a_i+1)q^{-d_i}) \subseteq [0,1)^s$$

in base q with $a_i, d_i \in \mathbb{Z}$, $d_i \geq 0$, and $0 \leq a_i < q^{d_i}$ for $1 \leq i \leq s$ and with $\lambda_s(J) = q^{t-m}$, that is, with $\sum_{i=1}^{s} d_i = m - t$. Then (4.64) shows that $\mathbf{x}_n \in J$ if and only if

$$T_\infty(C^{(i)}\mathbf{n}) \in [a_i q^{-d_i}, (a_i+1)q^{-d_i}) \qquad \text{for } 1 \leq i \leq s.$$

Since $C^{(i)}\mathbf{n} \in \mathbb{F}_q^\omega$ and $\psi(0) = 0$, the q-adic expansion of $T_\infty(C^{(i)}\mathbf{n})$ in (4.61) is finite, and so the condition above means that for $1 \leq i \leq s$ the first d_i q-adic digits of $T_\infty(C^{(i)}\mathbf{n})$ and $a_i q^{-d_i}$ agree. For each $j \geq 1$, the jth coordinate of $C^{(i)}\mathbf{n}$ is the "dot product" of the jth row $\mathbf{c}_j^{(i)}$ of $C^{(i)}$ with $\mathbf{n}$. Now in the given range $kq^m \leq n < (k+1)q^m$, the coordinates $\eta(z_r(n))$ of $\mathbf{n}$ are fixed for $r > m$, and so the jth coordinate of $C^{(i)}\mathbf{n}$ can be written as

$$\mathbf{c}_{j,m}^{(i)} \cdot (\eta(z_1(n)), \ldots, \eta(z_m(n))) + b_{k,m}^{(i)}$$

with $\mathbf{c}_{j,m}^{(i)}$ being the jth row vector of the submatrix $C_m^{(i)}$ of $C^{(i)}$ and with $b_{k,m}^{(i)} \in \mathbb{F}_q$ depending only on k, m, and $C^{(i)}$, but not on n. Thus, with

$$\mathbf{v} = (\eta(z_1(n)), \ldots, \eta(z_m(n)))^\top \in \mathbb{F}_q^m,$$

the condition that for $1 \leq i \leq s$ the first d_i q-adic digits of $T_\infty(C^{(i)}\mathbf{n})$ and $a_i q^{-d_i}$ agree is equivalent to $C_{m-t,m}\mathbf{v} = \mathbf{b}$ for some column vector $\mathbf{b} \in \mathbb{F}_q^{m-t}$ independent of n, where $C_{m-t,m}$ is an $(m-t) \times m$ matrix over $\mathbb{F}_q$ whose row vectors are the $\mathbf{c}_{j,m}^{(i)}$ with $1 \leq j \leq d_i$, $1 \leq i \leq s$. The definition of t in the theorem implies that $\varrho(C_m^{(1)}, \ldots, C_m^{(s)}) \geq m - t$, and so by the definition of the figure of merit the system $\{\mathbf{c}_{j,m}^{(i)} \in \mathbb{F}_q^m : 1 \leq j \leq m,\ 1 \leq i \leq s\}$ is an $(m-t, m, s)$-system over $\mathbb{F}_q$. Therefore the system $\{\mathbf{c}_{j,m}^{(i)} \in \mathbb{F}_q^m : 1 \leq j \leq d_i,\ 1 \leq i \leq s\}$ is linearly independent over $\mathbb{F}_q$, and so the equation $C_{m-t,m}\mathbf{v} = \mathbf{b}$ has exactly q^t solutions $\mathbf{v} \in \mathbb{F}_q^m$. Since η is a bijection, this yields exactly q^t integers n with $kq^m \leq n < (k+1)q^m$ such that $\mathbf{x}_n \in J$, and the desired net property is established. □

Example 4.4.64 We return to Example 4.4.62 and consider the generating matrix $C^{(1)}$ there. For every integer $m \geq 1$, the upper left $m \times m$ submatrix $C_m^{(1)}$ of $C^{(1)}$ is the $m \times m$ identity matrix over $\mathbb{F}_q$. It is trivial that $\varrho(C_m^{(1)}) = m$, and so the value of t in Theorem 4.4.63 is $t = 0$. This is consistent with the conclusion in Example 4.4.62.

We conclude this subsection by establishing an analog of Proposition 4.4.57 for digital (t, s)-sequences over $\mathbb{F}_q$.

Proposition 4.4.65 *Let q be a prime power and let $s \geq 1$ and $t \geq 0$ be integers. If a digital (t,s)-sequence over $\mathbb{F}_q$ is given, then for every integer $m \geq \max(1,t)$ we can construct a digital $(t,m,s+1)$-net over $\mathbb{F}_q$.*

Proof. We fix the integer $m \geq \max(1,t)$. As we can see from the proof of Theorem 4.4.63, the property of being a digital (t,s)-sequence over $\mathbb{F}_q$ depends only on its generating matrices and not on the bijections ψ and η. Thus, we are free to choose $\psi : \mathbb{F}_q \to Z_q$ and $\eta : Z_q \to \mathbb{F}_q$ as inverse maps of each other. Now let $(\mathbf{x}_n)_{n=0}^{\infty}$ be a digital (t,s)-sequence over $\mathbb{F}_q$ with $\infty \times \infty$ generating matrices $C^{(1)}, \ldots, C^{(s)}$ over $\mathbb{F}_q$. Then by Proposition 4.4.57, the points

$$\mathbf{y}_n = (nq^{-m}, \mathbf{x}_n) \in [0,1)^{s+1} \qquad \text{for } n = 0,1,\ldots,q^m - 1$$

form a $(t,m,s+1)$-net in base q. The definition of such a net and also the proof of Proposition 4.4.57 show that for the verification of this net property, we need to consider only $(s+1)$-dimensional elementary intervals J in base q as in the proof of Proposition 4.4.57 with $\sum_{i=1}^{s+1} d_i = m - t$ (where we write q for b), and so in particular with $d_i \leq m$ for $1 \leq i \leq s+1$. For checking whether $\mathbf{x}_n \in J$, only the first m q-adic digits of each coordinate matter. Thus, if $\mathbf{p}_n \in [0,1)^s$ is the point that is obtained by truncating each coordinate of $\mathbf{x}_n$ after the first m q-adic digits, then the points

$$\mathbf{w}_n = (nq^{-m}, \mathbf{p}_n) \in [0,1)^{s+1} \qquad \text{for } n = 0,1,\ldots,q^m - 1$$

form again a $(t,m,s+1)$-net in base q.

Now we construct $m \times m$ generating matrices $D^{(1)}, \ldots, D^{(s+1)}$ over $\mathbb{F}_q$ as follows. For $i = 1$ we let $D^{(1)} = (c_{ij})_{1 \leq i,j \leq m}$ be the antidiagonal matrix with $c_{ij} = 1$ if $i + j = m + 1$ and $c_{ij} = 0$ otherwise. For $2 \leq i \leq s+1$ we put $D^{(i)} = C_m^{(i-1)}$ with the notation in Theorem 4.4.63. If we write the column vector $\mathbf{v} \in \mathbb{F}_q^m$ in (4.49) in the form

$$\mathbf{v} = (\eta(z_1(n)), \ldots, \eta(z_m(n)))^{\top}$$

with $z_1(n), \ldots, z_m(n)$ running independently through Z_q and representing n for $0 \leq n \leq q^m - 1$ via (4.62), then it is straightforward to verify that the points $\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_{q^m-1}$ form a digital $(t,m,s+1)$-net over $\mathbb{F}_q$ with generating matrices $D^{(1)}, \ldots, D^{(s+1)}$. □

4.4.5 A construction of (t,s)-sequences

We describe a construction that, for every dimension $s \geq 1$ and every prime power q, produces a digital (t,s)-sequence over $\mathbb{F}_q$ for some value of the quality parameter t. This construction is due to Niederreiter [131] and it was historically the first construction achieving this task. We present a special case of this construction that is sufficient for our purposes and we refer to the paper [131] for the general case.

Given a dimension $s \geq 1$ and a prime power q, the basic ingredients of the construction are s distinct monic polynomials $p_1, \ldots, p_s \in \mathbb{F}_q[x]$ that are irreducible over the finite field $\mathbb{F}_q$. We put $l_i = \deg(p_i)$ for $1 \leq i \leq s$. The essential technical device is the expansion of rational functions over $\mathbb{F}_q$ into formal Laurent series over $\mathbb{F}_q$ that we already employed in Subsection 4.4.3 in the context of polynomial lattice point sets. Concretely, for $1 \leq i \leq s$ and for integers j and k with $j \geq 1$ and $0 \leq k < l_i$, we consider the rational function $x^k/p_i(x)^j \in \mathbb{F}_q(x)$. Since $\deg(x^k) = k < l_i \leq jl_i = \deg(p_i(x)^j)$, its formal Laurent series expansion has the form

$$\frac{x^k}{p_i(x)^j} = \sum_{r=1}^{\infty} e^{(i)}(j,k,r)x^{-r} \in \mathbb{F}_q((x^{-1})) \tag{4.65}$$

with coefficients $e^{(i)}(j,k,r) \in \mathbb{F}_q$. From these coefficients we derive the entries of the $\infty \times \infty$ generating matrices $C^{(1)}, \ldots, C^{(s)}$ over $\mathbb{F}_q$. We write $C^{(i)} = (c_{jr}^{(i)})_{j \geq 1,\, r \geq 1}$ for $1 \leq i \leq s$. For given i, j, r with $1 \leq i \leq s$, $j \geq 1$, and $r \geq 1$, we determine the entry $c_{jr}^{(i)}$ as follows. We express the integer $j-1$ uniquely as $j-1 = Q(i,j)l_i + k(i,j)$ with integers $Q(i,j)$ and $k(i,j)$ satisfying $Q(i,j) \geq 0$ and $0 \leq k(i,j) < l_i$. Then we put

$$c_{jr}^{(i)} = e^{(i)}(Q(i,j)+1, k(i,j), r) \in \mathbb{F}_q \qquad \text{for } 1 \leq i \leq s,\ j \geq 1,\ r \geq 1. \tag{4.66}$$

There is a condition that we need to check, namely that each column of $C^{(i)}$, $i = 1, \ldots, s$, contains only finitely many nonzero entries. If we fix i and an integer $r \geq 1$, then the entries in the rth column of $C^{(i)}$ are the elements $c_{jr}^{(i)}$ for $j = 1, 2, \ldots$. With ν denoting again the degree map on $\mathbb{F}_q((x^{-1}))$ introduced in Subsection 4.4.3, we obtain

$$\nu(x^k/p_i(x)^j) = \deg(x^k) - \deg(p_i(x)^j) = k - jl_i < l_i - jl_i,$$

and so $\nu(x^k/p_i(x)^j) \to -\infty$ as $j \to \infty$. This means that $e^{(i)}(j,k,r) = 0$ for all sufficiently large j, and so (4.66) implies that $c_{jr}^{(i)} = 0$ for all sufficiently large j.

The resulting generating matrices $C^{(1)}, \ldots, C^{(s)}$ yield an s-dimensional digital sequence over $\mathbb{F}_q$, called a *Niederreiter sequence* with generating polynomials $p_1, \ldots, p_s \in \mathbb{F}_q[x]$. It is remarkable that this construction always produces a digital (t,s)-sequence over $\mathbb{F}_q$ with a known value of t.

Theorem 4.4.66 *Let q be a prime power, let $s \geq 1$ be an integer, and let $p_1, \ldots, p_s \in \mathbb{F}_q[x]$ be distinct monic irreducible polynomials over $\mathbb{F}_q$. Then a Niederreiter sequence with generating polynomials $p_1, \ldots, p_s$ is a digital (t,s)-sequence over $\mathbb{F}_q$ with*

$$t = \sum_{i=1}^{s} (\deg(p_i) - 1).$$

Proof. With the given value of t, it suffices to show by Theorem 4.4.63 that $\varrho(C_m^{(1)}, \dots, C_m^{(s)}) \geq m - t$ for all $m \in \mathbb{N}$. This is trivial for $m \leq t$, and so we can assume that $m > t$. If

$$\mathbf{c}_{j,m}^{(i)} = (c_{j1}^{(i)}, \dots, c_{jm}^{(i)}) \in \mathbb{F}_q^m$$

denotes the jth row vector of $C_m^{(i)}$, then we have to verify that the system $\{\mathbf{c}_{j,m}^{(i)} : 1 \leq j \leq d_i,\ 1 \leq i \leq s\}$ is linearly independent over $\mathbb{F}_q$ for any integers $d_1, \dots, d_s \geq 0$ with $\sum_{i=1}^{s} d_i = m - t$. Let $n_i = \lceil d_i/l_i \rceil l_i$ for $1 \leq i \leq s$ be the least multiple of l_i that is greater than or equal to d_i. Then we prove even more, namely that the system $\{\mathbf{c}_{j,m}^{(i)} : 1 \leq j \leq n_i,\ 1 \leq i \leq s\}$ is linearly independent over $\mathbb{F}_q$.

Thus, suppose that

$$\sum_{i=1}^{s} \sum_{j=1}^{n_i} b_{i,j} \mathbf{c}_{j,m}^{(i)} = \mathbf{0} \in \mathbb{F}_q^m$$

for some $b_{i,j} \in \mathbb{F}_q$. A comparison of coordinates yields

$$\sum_{i=1}^{s} \sum_{j=1}^{n_i} b_{i,j} c_{jr}^{(i)} = 0 \qquad \text{for } 1 \leq r \leq m. \tag{4.67}$$

Consider the rational function

$$R := \sum_{i=1}^{s} \sum_{j=1}^{n_i} b_{i,j} \frac{x^{k(i,j)}}{p_i(x)^{Q(i,j)+1}} = \sum_{r=1}^{\infty} \Big(\sum_{i=1}^{s} \sum_{j=1}^{n_i} b_{i,j} c_{jr}^{(i)} \Big) x^{-r},$$

where we used (4.65) and (4.66) in the second identity. Note that $\nu(R) < -m$ by (4.67). For $1 \leq i \leq s$ we can write

$$\begin{aligned} \sum_{j=1}^{n_i} b_{i,j} \frac{x^{k(i,j)}}{p_i(x)^{Q(i,j)+1}} &= \sum_{h=1}^{n_i/l_i} \sum_{j=(h-1)l_i+1}^{hl_i} \frac{b_{i,j} x^{j-1-(h-1)l_i}}{p_i(x)^h} \\ &= \sum_{h=1}^{n_i/l_i} \frac{1}{p_i(x)^h} \sum_{k=0}^{l_i-1} b_{i,(h-1)l_i+k+1} x^k = \sum_{h=1}^{n_i/l_i} \frac{f_{ih}(x)}{p_i(x)^h}, \end{aligned}$$

where

$$f_{ih}(x) = \sum_{k=0}^{l_i-1} b_{i,(h-1)l_i+k+1} x^k \in \mathbb{F}_q[x] \qquad \text{for } 1 \leq h \leq n_i/l_i. \tag{4.68}$$

If we put $g(x) = \prod_{i=1}^{s} p_i(x)^{n_i/l_i}$, then Rg is a polynomial. On the other hand,

$$\begin{aligned} \nu(Rg) < -m + \deg(g) &= -m + \sum_{i=1}^{s} n_i \leq -m + \sum_{i=1}^{s} (d_i + l_i - 1) \\ &= -m + m - t + \sum_{i=1}^{s} (l_i - 1) = 0. \end{aligned}$$

This is possible only if $Rg = 0$, and so $R = 0$. Hence we obtain

$$\sum_{i=1}^{s} \sum_{h=1}^{n_i/l_i} \frac{f_{ih}}{p_i^h} = R = 0. \tag{4.69}$$

If we can show that $f_{ih} = 0 \in \mathbb{F}_q[x]$ for $1 \leq h \leq n_i/l_i$, $1 \leq i \leq s$, then all $b_{i,j} = 0$ and we are done. There is nothing to prove if $n_i = 0$, and so we consider only those i with $n_i \geq 1$, that is, with $n_i \geq l_i$. We multiply (4.69) by g and obtain the polynomial identity

$$\sum_{i=1}^{s} \Big(\sum_{h=1}^{n_i/l_i} f_{ih} p_i^{(n_i/l_i)-h} \Big) \prod_{\substack{a=1 \\ a \neq i}}^{s} p_a^{n_a/l_a} = 0. \tag{4.70}$$

Now we fix an integer i with $n_i \geq l_i$. Then p_i divides the right-hand side of (4.70), and so p_i must divide the left-hand side. It follows that p_i divides

$$\Big(\sum_{h=1}^{n_i/l_i} f_{ih} p_i^{(n_i/l_i)-h} \Big) \prod_{\substack{a=1 \\ a \neq i}}^{s} p_a^{n_a/l_a}.$$

Since p_i does not divide the last product, p_i must divide the last sum by Proposition 1.4.17(ii), and so p_i divides f_{ih} for $h = n_i/l_i$. But $\deg(f_{ih}) < l_i = \deg(p_i)$ by (4.68), and so $f_{ih} = 0$ for $h = n_i/l_i$. This means that the terms in (4.69) corresponding to $h = n_i/l_i$ with $n_i \geq l_i$ drop out. By repeating this argument sufficiently often, we arrive at the desired conclusion that $f_{ih} = 0$ for $1 \leq h \leq n_i/l_i$, $1 \leq i \leq s$. □

Remark 4.4.67 The result of Theorem 4.4.66 is best possible since it was proved by Dick and Niederreiter [38] that a Niederreiter sequence with generating polynomials $p_1, \ldots, p_s \in \mathbb{F}_q[x]$ cannot be a digital (v, s)-sequence over $\mathbb{F}_q$ for an integer $v < \sum_{i=1}^{s}(\deg(p_i) - 1)$.

Example 4.4.68 For every prime power q, there are exactly q distinct monic *linear* polynomials over $\mathbb{F}_q$ which are of course automatically irreducible over $\mathbb{F}_q$. Hence for every dimension s with $1 \leq s \leq q$, we can choose distinct monic linear polynomials $p_1, \ldots, p_s \in \mathbb{F}_q[x]$. Then by Theorem 4.4.66, a Niederreiter sequence with these generating polynomials $p_1, \ldots, p_s$ is a digital $(0, s)$-sequence over $\mathbb{F}_q$. This result is noteworthy since Theorem 4.4.58 says that $s \leq q$ is a necessary condition for the existence of a $(0, s)$-sequence in base q. Therefore we get the elegant statement that for a prime power q, a $(0, s)$-sequence in base q exists if and only if $s \leq q$, and if $s \leq q$ holds, then we can even construct a digital $(0, s)$-sequence over $\mathbb{F}_q$. For the Niederreiter sequences that are digital $(0, s)$-sequences over $\mathbb{F}_q$, the generating matrices can be written down in a nice explicit form (see [134, Remark 4.52]). With the approach via these explicit generating matrices, these digital $(0, s)$-sequences over $\mathbb{F}_q$ were constructed earlier by Faure [49] for prime numbers q and by Niederreiter [130] for arbitrary prime powers q.

Example 4.4.69 If we combine the construction in Example 4.4.68 with Proposition 4.4.65, then for every prime power q, for every dimension s with $2 \leq s \leq q+1$, and for every integer $m \geq 1$ we obtain a digital $(0, m, s)$-net over $\mathbb{F}_q$. Such digital nets were also constructed in Theorem 4.4.53 by a different method. The existence of these digital nets over $\mathbb{F}_q$ for $s = 1$ is known from Example 4.4.16. As soon as $m \geq 2$, the condition $s \leq q + 1$ is best possible in the light of Theorem 4.4.11.

Remark 4.4.70 Given a prime power q and a dimension $s \geq 1$, the problem of minimizing the value of the quality parameter t in Theorem 4.4.66 is easy to solve. We just have to choose for $p_1, \ldots, p_s$ distinct monic irreducible polynomials of least degrees. More formally, we list all monic irreducible polynomials over $\mathbb{F}_q$ (there are infinitely many of them by Proposition 1.4.43) in a sequence according to nondecreasing degrees and then we let $p_1, \ldots, p_s$ be the first s terms of this sequence. With such a choice for $p_1, \ldots, p_s$, we put

$$P_q(s) = \sum_{i=1}^{s} (\deg(p_i) - 1).$$

The polynomials $p_1, \ldots, p_s$ are not uniquely determined since we are not saying anything about the order in which monic irreducible polynomials over $\mathbb{F}_q$ of the same degree are listed, but the number $P_q(s)$ is well defined. For instance, for $q = 2$ and $s = 6$ we can take $p_1(x) = x$, $p_2(x) = x+1$, $p_3(x) = x^2+x+1$, $p_4(x) = x^3+x+1$, $p_5(x) = x^3 + x^2 + 1$, $p_6(x) = x^4 + x + 1$, and so $P_2(6) = 8$. For every prime power q and every integer $s \geq 1$, we get a digital $(P_q(s), s)$-sequence over $\mathbb{F}_q$ by Theorem 4.4.66. For $1 \leq s \leq q$ we obtain $P_q(s) = 0$ by Example 4.4.68, and for $s > q$ we know the bound

$$P_q(s) < s(\log_q s + \log_q \log_q s + 1),$$

where $\log_q$ denotes the logarithm to the base q (see [134, Theorem 4.54] for a proof of this bound).

Theorem 4.4.66 has an appealing consequence that we already advertised at the beginning of this subsection, namely that for every prime power q and every integer $s \geq 1$, there exists a digital (t, s)-sequence over $\mathbb{F}_q$ for some value of t. From the practical point of view, there is great interest in the least value of t that can be achieved by any kind of construction with the digital method, and this leads to the following definition.

Definition 4.4.71 For every prime power q and every integer $s \geq 1$, let $d_q(s)$ be the least value of t for which there exists a digital (t, s)-sequence over $\mathbb{F}_q$.

Example 4.4.68 shows that $d_q(s) = 0$ for $1 \leq s \leq q$ and Theorem 4.4.58 implies that $d_q(s) \geq 1$ for $s \geq q+1$. For a fixed prime power q, the following theorem says that $d_q(s)$ grows at least linearly as a function of s as s tends to ∞. The proof of this result uses concepts and facts from coding theory (see Chapter 3).

Theorem 4.4.72 *The lower bound*

$$d_q(s) \geq \frac{s}{q} - \log_q \frac{(q-1)s+q+1}{2}$$

holds for all prime powers q and all integers $s \geq 1$.

Proof. We fix q and s and observe that for $t = d_q(s)$ there exists a digital (t,s)-sequence over $\mathbb{F}_q$. Then Proposition 4.4.65 shows that for every integer $m > t$ there is a digital $(t,m,s+1)$-net over $\mathbb{F}_q$. We put $h = \lfloor (q-1)s/q \rfloor + 1$ and consider $m = t+h$. If $s+1 \leq m$, then

$$d_q(s) = t = m - h \geq s + 1 - (q-1)s/q - 1 = s/q$$

and we are done. Thus, we can assume that $s+1 > m$. By Theorem 4.4.19, our digital $(t,m,s+1)$-net over $\mathbb{F}_q$ with $m = t+h$ yields an $(h,m,s+1)$-system $\{\mathbf{c}_j^{(i)} \in \mathbb{F}_q^m : 1 \leq j \leq m,\ 1 \leq i \leq s+1\}$ over $\mathbb{F}_q$. We apply the definition of an $(h,m,s+1)$-system over $\mathbb{F}_q$ (see Definition 4.4.18) only to the vectors $\mathbf{c}_j^{(i)}$ with $j=1$. Then we infer that any h of the vectors $\mathbf{c}_1^{(i)}$, $1 \leq i \leq s+1$, are linearly independent over $\mathbb{F}_q$.

Now we set up the $m \times (s+1)$ matrix H over $\mathbb{F}_q$ with the column vectors $\mathbf{c}_1^{(1)}, \ldots, \mathbf{c}_1^{(s+1)}$. Then we consider the subspace $\{\mathbf{v} \in \mathbb{F}_q^{s+1} : \mathbf{v}H^\top = \mathbf{0} \in \mathbb{F}_q^m\}$ of $\mathbb{F}_q^{s+1}$. This is a linear code over $\mathbb{F}_q$ of length $s+1$, of dimension at least $s+1-m$, and with minimum distance at least $h+1$ (see the proof of Theorem 3.2.44). By passing to an $(s+1-m)$-dimensional subspace of this linear code, we arrive at a linear $[s+1, s+1-m]$ code C over $\mathbb{F}_q$ with minimum distance $d(C) \geq h+1$. We apply the Plotkin bound in Theorem 3.4.19 to the linear code C, and this yields

$$h + 1 \leq d(C) \leq \frac{(s+1)(q-1)q^{s-m}}{q^{s+1-m}-1}.$$

A straightforward manipulation using $m = t+h$ shows that

$$q^{t+h-s} \geq \frac{(h+1)q - (s+1)(q-1)}{h+1},$$

and by taking logarithms to the base q we obtain

$$d_q(s) = t \geq s - h + \log_q \Big(q - \frac{(s+1)(q-1)}{h+1} \Big).$$

Next we note that $h = \lfloor (q-1)s/q \rfloor + 1 \leq (q-1)s/q + 1$, hence

$$d_q(s) \geq \frac{s}{q} - 1 + \log_q \Big(q - \frac{(s+1)(q-1)}{h+1} \Big).$$

Furthermore $h \geq s - s/q + 1/q$, therefore

$$q - \frac{(s+1)(q-1)}{h+1} \geq \frac{2q}{(q-1)s + q + 1},$$

and the desired lower bound on $d_q(s)$ follows. □

With $P_q(s)$ as in Remark 4.4.70, we obviously have $d_q(s) \leq P_q(s)$ for all prime powers q and all integers $s \geq 1$. The upper bound on $P_q(s)$ in Remark 4.4.70 implies that, for fixed q, the quantity $d_q(s)$ is at most of the order of magnitude $s \log s$. On the other hand, Theorem 4.4.72 shows that $d_q(s)$ is at least of the order of magnitude s. Actually, $d_q(s)$ has the order of magnitude s as a function of s as s tends to ∞, but this can be proved only by deeper methods which are beyond the scope of this book (see Section 4.5 for a sketch of these methods).

4.5 A glimpse of advanced topics

There are various notions of discrepancy besides the extreme discrepancy and the star discrepancy. For a point set $\mathcal{P}$ consisting of N points in $[0,1)^s$, we introduce a function of $\mathbf{u} = (u_1, \ldots, u_s) \in [0,1]^s$ by

$$R_{\mathcal{P}}(\mathbf{u}) = N^{-1} A \Big(\prod_{i=1}^{s} [0, u_i); \mathcal{P} \Big) - u_1 \cdots u_s.$$

Then the star discrepancy $D_N^*(\mathcal{P})$ of $\mathcal{P}$ is the supremum norm of the function $R_{\mathcal{P}}$ on $[0,1]^s$. For a real number $p \geq 1$, the L^p norm of the function $R_{\mathcal{P}}$ on $[0,1]^s$ is called the *L^p discrepancy* of $\mathcal{P}$. The case $p = 2$ has received special attention. The lower bound of Roth [176] in (4.19) is actually a lower bound on the L^2 discrepancy of $\mathcal{P}$. Other concepts of discrepancy are obtained by extending the supremum in (4.18) not only over subintervals of $[0,1]^s$, but over more general sets such as convex subsets of $[0,1]^s$. This leads then also to error bounds for quasi-Monte Carlo integration for wider classes of integration domains, for instance for convex integration domains or for Jordan-measurable integration domains. We refer to [127, Sections 2 and 3] and [134, Chapter 2] for discussions of these topics.

Quasi-Monte Carlo methods can be applied not only to numerical integration, but to a variety of other tasks in computational mathematics. We mention the numerical solution of integral equations, of integro-differential equations, and of linear partial differential equations. Perhaps more surprising are applications to approximation theory and to the computation of eigenvalues of matrices. We refer to [41, Chapter 3] for some of these applications. A theory that is quite well developed is that of the quasi-Monte Carlo method for the computation of maxima and minima of real-valued functions, also called quasirandom search (see [134, Chapter 6]).

Much more can be said about lattice rules and there is even a book devoted entirely to lattice rules (see [189]). With every s-dimensional lattice L we can associate an $s \times s$ *generator matrix* B whose row vectors $\mathbf{b}_1, \ldots, \mathbf{b}_s$ form a basis of L in the sense of (4.45). If L is an integration lattice, then the absolute value of the determinant of B is equal to $1/N$, where N is the number of points in the lattice point set corresponding to L. Furthermore, a generator matrix of the dual lattice $L^\perp$ of the integration lattice L is given by $(B^\top)^{-1}$, that is, the inverse of the transpose of B. We proved existence theorems for good lattice rules of rank 1 in Subsection 4.3.1, and there are also existence theorems for good lattice rules of higher rank (see [134, Section 5.4] and [189]). An important development for the practical computation of good lattice points is the fast CBC algorithm of Nuyens and Cools [154] which is based on fast Fourier transform techniques.

There are more constructions of nets and (t, s)-sequences than those presented in Subsections 4.4.3 and 4.4.5. A detailed expository account of further constructions can be found in the book of Dick and Pillichshammer [39]. A fascinating issue is that of the exact order of magnitude of the quantity $d_q(s)$ introduced in Definition 4.4.71, where we fix the prime power q and consider $d_q(s)$ as a function of the dimension s. A lower bound on $d_q(s)$ of the order of magnitude s was established in Theorem 4.4.72. This is actually the exact order of magnitude of $d_q(s)$, but to prove this we need also an upper bound on $d_q(s)$ of the same order of magnitude. This means that for every $s \geq 1$ we have to construct a digital (t, s)-sequence over $\mathbb{F}_q$ with t growing linearly as a function of s. All the known constructions that achieve this rate of growth use the theory of global function fields outlined in Section 3.6.

The family of *Niederreiter-Xing sequences* was the first family of constructions that produced an upper bound on $d_q(s)$ of the order of magnitude s. The main papers here are [146] and [203], and accounts of these constructions are also given in the books [39, Chapter 8] and [147, Chapter 8]. There is a minor technical issue that arises in these and several other constructions, namely that it cannot be guaranteed any more that the $\infty \times \infty$ generating matrices $C^{(1)}, \ldots, C^{(s)}$ over $\mathbb{F}_q$ have the property that each column of each generating matrix contains only finitely many nonzero entries. This situation is remedied by slightly modifying the definition of a (t, s)-sequence in base b. Let $[\mathbf{x}]_{b,m} \in [0, 1)^s$ denote the point that is obtained by the coordinatewise m-digit truncation in base b of the point $\mathbf{x} \in [0, 1]^s$. Then we say that for integers $b \geq 2$, $s \geq 1$, and $t \geq 0$, a sequence $(\mathbf{x}_n)_{n=0}^{\infty}$ of points in the closed s-dimensional unit cube $[0, 1]^s$ is a (t, s)-sequence in base b *in the broad sense* if, for all integers $k \geq 0$ and $m > t$, the points $[\mathbf{x}_n]_{b,m}$ with $kb^m \leq n < (k + 1)b^m$ form a (t, m, s)-net in base b. A (t, s)-sequence in base b satisfying the original Definition

4.4.55 is then called a (t,s)-sequence in base b *in the narrow sense.* Analogously, we speak of digital (t,s)-sequences over $\mathbb{F}_q$ in the broad sense and in the narrow sense. The main results on (t,s)-sequences in base b in the narrow sense, such as the discrepancy bound in Theorem 4.4.59 with an obvious notion of star discrepancy for sequences of points in $[0,1]^s$, hold just as well for (t,s)-sequences in base b in the broad sense.

Now we sketch a construction from the family of Niederreiter-Xing sequences, namely the construction in [146] using rational places. For a given prime power q and a given integer $s \geq 1$, we choose a global function field $F/\mathbb{F}_q$ containing at least $s+1$ rational places. Let $P_\infty, P_1, \dots, P_s$ be $s+1$ distinct rational places of F. Furthermore, we choose a divisor $D \geq 0$ of F with $\dim(\mathcal{L}(D+jP_i)) = j+1$ for $1 \leq i \leq s$ and all integers $j \geq 0$. Then for each $i = 1, \dots, s$ and $j \geq 1$, there is an element

$$f_j^{(i)} \in \mathcal{L}(D+jP_i) \setminus \mathcal{L}(D+(j-1)P_i).$$

Next we pick an element $y \in F$ with $\nu_{P_\infty}(y) = 1$. Recall that in the rational function field $\mathbb{F}_q(x)$ there is an expansion into formal Laurent series in terms of powers of x (see Subsection 4.4.3). There is an analogous expansion in the global function field F in terms of powers of y. For the elements $f_j^{(i)}$ above, these expansions can be written in the form

$$f_j^{(i)} = y^{-v} \sum_{r=0}^{\infty} b_{j,r}^{(i)}\, y^r \qquad \text{for } 1 \leq i \leq s \text{ and } j \geq 1,$$

where all coefficients $b_{j,r}^{(i)} \in \mathbb{F}_q$ and where the integer $v \geq 0$ is the coefficient of P_∞ in the representation of the divisor D as a formal linear combination of places. For $i = 1, \dots, s$, we now set up the $\infty \times \infty$ generating matrix $C^{(i)} = (c_{j,r}^{(i)})_{j\geq 1,\, r\geq 0}$ over $\mathbb{F}_q$ by putting

$$c_{j,r}^{(i)} = \begin{cases} b_{j,r}^{(i)} & \text{for } j \geq 1 \text{ and } 0 \leq r \leq v-1, \\ b_{j,r+1}^{(i)} & \text{for } j \geq 1 \text{ and } r \geq v. \end{cases}$$

These generating matrices $C^{(1)}, \dots, C^{(s)}$ yield a digital (t,s)-sequence over $\mathbb{F}_q$ in the broad sense with $t = g$, the genus of the global function field F.

The problem of optimizing this construction of Niederreiter-Xing sequences leads naturally to the quantity $V_q(s)$ which, for every prime power q and every integer $s \geq 1$, is defined as the least integer $g \geq 0$ for which there exists a global function field $F/\mathbb{F}_q$ of genus g containing at least $s+1$ rational places. Then obviously $d_q(s) \leq V_q(s)$, where we include digital (t,s)-sequences over $\mathbb{F}_q$ in the broad sense in the definition of $d_q(s)$. The quantity $V_q(s)$ can be bounded by means of the so-called class field theory of global function fields. As a consequence, for every prime power q we get the bound

$$d_q(s) \leq \frac{cs}{\log q} + 1 \qquad \text{for all } s \geq 1,$$

where $c > 0$ is an absolute constant. This bound settles the problem of the exact order of magnitude of $d_q(s)$ as a function of s. Various other bounds on $d_q(s)$ can be derived from Niederreiter-Xing sequences. For instance, the agreeable bound $d_q(s) \leq 5s$ for all prime powers q and all $s \geq 1$ is an immediate consequence of [203, Theorem 3]. For special values of q some better bounds are possible; for instance if q is a square, then

$$d_q(s) \leq \frac{ps}{q^{1/2}-1} \qquad \text{for all } s \geq 1,$$

where p is the unique prime factor of q. Detailed information on these and other bounds for $d_q(s)$ is available in [147, Sections 8.3 and 8.4].

There are other constructions based on global function fields that yield digital (g, s)-sequences over $\mathbb{F}_q$, just like the Niederreiter-Xing sequence described above. A relatively simple one is that of Hofer and Niederreiter [66] which does not require the auxiliary divisor D in the Niederreiter-Xing construction. The construction by Niederreiter and Yeo [149] stands out because it is the only known construction of (t, s)-sequences that operates for every dimension s and that is not based on the digital method. In fact, this construction is a relative of the construction of Halton sequences in Subsection 4.2.2, in the sense that it uses a sort of radical-inverse function in the context of global function fields.

We presented the details and the proof for only one construction of (t, s)-sequences that works for every dimension s, namely the construction of Niederreiter sequences in Subsection 4.4.5. This construction can be applied only for prime-power bases q. What can be said about bases b that are not prime powers? The approach for such a base b is to use again a digital method, but instead of a finite field one employs a finite commutative ring R with identity and of cardinality b as the underlying algebraic structure. A convenient choice is obtained by writing $b = \prod_{h=1}^{r} q_h$ as a product of pairwise coprime prime powers $q_1, \ldots, q_r$ and by letting R be the direct product $R = \prod_{h=1}^{r} \mathbb{F}_{q_h}$ of finite fields. Then constructions that work over finite fields can be extended to R by a direct-product procedure.

A quantity that is more general than $d_q(s)$ is the number $t_b(s)$ which is defined, for all integers $b \geq 2$ and $s \geq 1$, as the least value of t for which there exists a (t, s)-sequence in base b (in the broad sense, say). It is obvious that for prime powers q the inequality $t_q(s) \leq d_q(s)$ holds for all $s \geq 1$. By using the direct-product procedure mentioned above, together with the Niederreiter-Xing construction, it was shown in [146] that for every base $b \geq 2$ the upper bound

$$t_b(s) \leq \frac{cs}{\log q(b)} + 1$$

is valid for all $s \geq 1$, where $c > 0$ is an absolute constant and where $q(b)$ is the smallest prime power in the factorization of b into pairwise coprime prime powers. On the other hand, there is a lower bound on $t_b(s)$ which, for fixed b, is also linear in s, and so $t_b(s)$ has the exact order of magnitude s just like $d_q(s)$. We refer to [147, Chapter 8] for more information on (t, s)-sequences in arbitrary bases b.

Exercises

4.1 Prove that a sequence $(x_n)_{n=1}^{\infty}$ of points in $[0, 1)$ is uniformly distributed if and only if (4.7) holds for every subinterval J of $[0, 1]$ with rational endpoints.

4.2 Prove that if finitely many terms of a sequence that is uniformly distributed modulo 1 are deleted or changed in an arbitrary manner, then the resulting sequence is still uniformly distributed modulo 1.

4.3 Prove that if the sequences $(x_n)_{n=1}^{\infty}$ and $(y_n)_{n=1}^{\infty}$ are uniformly distributed modulo 1, then the "mixed" sequence $x_1, y_1, x_2, y_2, \ldots, x_n, y_n, \ldots$ is uniformly distributed modulo 1.

4.4 Prove that the sequence $\frac{0}{1}, \frac{0}{2}, \frac{1}{2}, \frac{0}{3}, \frac{1}{3}, \frac{2}{3}, \ldots, \frac{0}{k}, \frac{1}{k}, \ldots, \frac{k-1}{k}, \ldots$ constructed in an obvious blockwise manner is uniformly distributed.

4.5 Let m be a nonzero integer and let c be a real number. Prove that if the sequence $(x_n)_{n=1}^{\infty}$ is uniformly distributed modulo 1, then so is the sequence $(mx_n + c)_{n=1}^{\infty}$.

4.6 Let $(x_n)_{n=1}^{\infty}$ and $(y_n)_{n=1}^{\infty}$ be two sequences of real numbers such that $\lim_{n\to\infty}(x_n - y_n) = c$ for some $c \in \mathbb{R}$. Prove that if $(x_n)_{n=1}^{\infty}$ is uniformly distributed modulo 1, then so is $(y_n)_{n=1}^{\infty}$. (Hint: use Theorem 4.1.9.)

4.7 Prove that if $(x_n)_{n=1}^{\infty}$ is uniformly distributed modulo 1 and the sequence $(y_n)_{n=1}^{\infty}$ of real numbers satisfies

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} |y_n| = 0,$$

then $(x_n + y_n)_{n=1}^{\infty}$ is uniformly distributed modulo 1. (Hint: use Theorem 4.1.9.)

4.8 For a point set $\mathcal{P}$ consisting of N points in $[0, 1)^s$ with $s \geq 1$, show that any given point in $[0, 1)^s$ can occur at most $\lfloor ND_N(\mathcal{P})\rfloor$ times in $\mathcal{P}$.

4.9 Let $x_1, \ldots, x_N \in [0, 1)$ be such that, for some constants $c > 0$ and $C_1 > 0$, the bound $\left|\sum_{n=1}^{N} e^{2\pi i h x_n}\right| \leq C_1 h^c$ holds for all integers h with $1 \leq h \leq N^{1/(c+1)}$. Prove that the point set $\mathcal{P}$ consisting of $x_1, \ldots, x_N$ satisfies $D_N(\mathcal{P}) \leq C_2 N^{-1/(c+1)}$ with a constant $C_2 > 0$ depending only on c and C_1.

4.10 Establish an s-dimensional version of Lemma 4.1.15 for every $s \geq 2$.

4.11 Theorem 4.1.21 implies that, with the notation in this theorem,

$$\Big|\frac{1}{N}\sum_{n=1}^{N} x_n - \frac{1}{2}\Big| \le D_N^*(\mathcal{P})$$

for all $x_1, \ldots, x_N \in [0,1)$. Prove that this is best possible in the sense that for every constant $c < 1$ there exist points $x_1, \ldots, x_N \in [0,1)$, where N may depend on c, such that

$$\Big|\frac{1}{N}\sum_{n=1}^{N} x_n - \frac{1}{2}\Big| \ge c D_N^*(\mathcal{P}).$$

(Hint: consider point sets of the form

$$\underbrace{0, \ldots, 0}_{m}, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-m}{N},$$

where $m \in \mathbb{N}$ with $m \le N$ is suitably chosen.)

4.12 Let $x_1, \ldots, x_N$ be real numbers and let D_N^* be the star discrepancy of their fractional parts. Prove that

$$\Big|\sum_{n=1}^{N} e^{2\pi \mathrm{i} x_n}\Big| \le 4ND_N^*.$$

4.13 For every point set $\mathcal{P}$ consisting of N points in $[0,1)^2$, prove that $D_N(\mathcal{P}) \le 4D_N^*(\mathcal{P})$. (Hint: express an arbitrary half-open subinterval of $[0,1)^2$ in terms of half-open subintervals anchored at the origin.)

4.14 Generalize the preceding exercise and show that $D_N(\mathcal{P}) \le 2^s D_N^*(\mathcal{P})$ for every point set $\mathcal{P}$ consisting of N points in $[0,1)^s$ with $s \ge 2$.

4.15 Prove in detail that if the real-valued function f on $[0,1]^s$ with $s \ge 2$ depends on fewer than s variables, then the variation $V^{(s)}(f)$ of f on $[0,1]^s$ in the sense of Vitali satisfies $V^{(s)}(f) = 0$.

4.16 With the notation for continued fractions in Subsection 4.2.1, prove that

$$p_k q_{k-2} - p_{k-2} q_k = (-1)^k a_k \qquad \text{for all } k \ge 0.$$

4.17 Prove that the star discrepancy of the point set $\mathcal{P}_{m,s}$ in Subsection 4.3.1 is equal to $1 - \big(1 - \frac{1}{2m}\big)^s$.

4.18 Let $m \geq 2$ and $s \geq 1$ be integers and let $\mathcal{P}$ be the point set consisting of the $N = m^s$ points

$$\left(\frac{k_1}{m}, \dots, \frac{k_s}{m}\right) \in [0,1)^s$$

with $k_1, \dots, k_s$ running independently through the integers $0, 1, \dots, m-1$. Prove that $D_N^*(\mathcal{P}) = 1 - \left(1 - \frac{1}{m}\right)^s$.

4.19 For every $\mathbf{g} \in \mathbb{Z}^s$ with $s \geq 2$ and every integer $N \geq 2$, define

$$\varrho(\mathbf{g}, N) = \min_{\mathbf{h}} r(\mathbf{h}),$$

where the minimum is extended over all nonzero $\mathbf{h} \in \mathbb{Z}^s$ with $\mathbf{h} \cdot \mathbf{g} \equiv 0 \pmod{N}$ and where $r(\mathbf{h})$ is given by (4.38). Prove that $1 \leq \varrho(\mathbf{g}, N) \leq N/2$.

4.20 Let $\mathbf{g}$, N, and $\varrho(\mathbf{g}, N)$ be as in the preceding exercise. Prove that

$$D_N^*(\mathcal{P}(\mathbf{g}, N)) \geq \frac{C_s}{\varrho(\mathbf{g}, N)}$$

with a constant $C_s > 0$ depending only on s. This shows that $\varrho(\mathbf{g}, N)$ must be large for a good lattice point $\mathbf{g}$ modulo N. (Hint: apply Theorem 4.1.41 with a function f of the form $f(\mathbf{u}) = \cos(2\pi \mathbf{h} \cdot \mathbf{u})$ for a suitable $\mathbf{h} \in \mathbb{Z}^s$.)

4.21 Let L be the two-dimensional lattice corresponding to $\mathcal{P}(\mathbf{g}, N)$ with $\mathbf{g} = (1, 5) \in \mathbb{Z}^2$ and $N = 8$. Determine the dual lattice $L^\perp$ explicitly and compute the number $\varrho(\mathbf{g}, N)$ in Exercise 4.19.

4.22 Consider the two-dimensional lattice point set consisting of the six points given by the fractional parts $\left\{j_1\left(\frac{1}{2}, 0\right) + j_2\left(\frac{1}{3}, \frac{1}{3}\right)\right\}$ with $j_1 \in \{0, 1\}$ and $j_2 \in \{0, 1, 2\}$. Determine the rank and the invariants of this lattice point set.

4.23 Prove that if there exists a (t, m, s)-net in base b, then for every integer $h \geq 1$ there exists a $(t+h, m+h, s)$-net in base b.

4.24 Prove that if there exists a digital (t, m, s)-net over $\mathbb{F}_q$, then for every integer $h \geq 1$ there exists a digital $(t+h, m+h, s)$-net over $\mathbb{F}_q$.

4.25 Prove that every (ht, hm, s)-net in base b with an integer $h \geq 1$ is a (t, m, s)-net in base b^h.

4.26 Prove that if there exists a (t_1, m_1, s_1)-net in base b and a (t_2, m_2, s_2)-net in base b, then there exists a $(t, m_1 + m_2, s_1 + s_2)$-net in base b with

$$t = \max(m_1 + t_2, m_2 + t_1).$$

(Hint: consider the $(s_1 + s_2)$-dimensional direct product of the s_1-dimensional net and the s_2-dimensional net.)

4.27 Prove that in Definition 4.4.18 we can replace the condition $\sum_{i=1}^{s} d_i = d$ by $\sum_{i=1}^{s} d_i \leq d$ and we still have the stated linear independence property.

4.28 Prove that if $C^{(1)}$ is the 4×4 identity matrix over $\mathbb{F}_2$ and the matrices $C^{(2)}$ and $C^{(3)}$ over $\mathbb{F}_2$ are given by

$$C^{(2)} = \begin{pmatrix} 0\,0\,0\,1 \\ 0\,0\,1\,0 \\ 0\,1\,0\,0 \\ 1\,0\,0\,0 \end{pmatrix}, \quad C^{(3)} = \begin{pmatrix} 1\,1\,1\,1 \\ 0\,1\,0\,1 \\ 0\,0\,1\,1 \\ 0\,0\,0\,1 \end{pmatrix},$$

then $C^{(1)}$, $C^{(2)}$, and $C^{(3)}$ are generating matrices of a digital $(0, 4, 3)$-net over $\mathbb{F}_2$.

4.29 There is an analog of the Korobov form (4.43) of lattice points in the context of polynomial lattice point sets. Choose $f, g \in \mathbb{F}_q[x]$ with $\deg(f) = m \geq 1$ and determine $\boldsymbol{g} = (g_1, \ldots, g_s) \in \mathbb{F}_q[x]^s_{<m}$ by $g_i \equiv g^{i-1} \pmod{f}$ and $\deg(g_i) < m$ for $1 \leq i \leq s$. Under the assumption that f is irreducible over $\mathbb{F}_q$, prove an analog of Theorem 4.4.46 for s-tuples $\boldsymbol{g} \in \mathbb{F}_q[x]^s_{<m}$ of Korobov form.

4.30 Show first that $f(x) = x^3 + 2x + 1 \in \mathbb{F}_3[x]$ is irreducible over $\mathbb{F}_3$ and then construct explicitly a quadruple $\boldsymbol{g} \in \mathbb{F}_3[x]^4_{<3}$ that yields a digital $(0, 3, 4)$-net over $\mathbb{F}_3$ according to Theorem 4.4.53.

4.31 Compute the quantity $P_q(s)$ in Remark 4.4.70 for $q = 2$ and $s = 10$.

4.32 Compute the quantity $P_q(s)$ in Remark 4.4.70 for $q = 3$ and $s = 8$.

Chapter 5

Pseudorandom Numbers

Random numbers, pseudo or true,
better look like out of the blue.
But what to do on a day
when the sky is black or gray,
there we don't have any clue.

5.1 General principles

5.1.1 Random number generation

We pointed out in Subsection 4.1.2 that the Monte Carlo method for numerical integration uses random samples, but we did not say anything about how to produce random samples. Maybe we were wise to keep quiet, because nobody really knows how to generate honest-to-goodness random samples in practice. On the other hand, the purely theoretical framework for random sampling is clear: we have a set with at least two elements, we are given a probability distribution (or a probability measure) on this set, and we want to pick elements from this set that fairly represent the probability distribution. But how do we decide whether the sampling is fair? Normally in mathematics there is a definition to which we refer for the verification of a property, but here no generally accepted definition of a fair random sample is codified, unless we deceive ourselves and tolerate tautologies like "a random sample is a collection of elements chosen at random".

Notwithstanding this conundrum about random samples, the procedure of random sampling is widely used in many walks of life. It is of course the mainstay of applied statistics where information on large populations is obtained by drawing and investigating a relatively small random sample. A typical example is an opinion poll in which only a tiny percentage of the populace is queried. This small group of people should form, as the statisticians say, a "representative sample" in terms of demographic categories like age, gender, and social class. Therefore it loosely captures the idea of a “fair random sample” mentioned earlier. However, when one considers how often pollsters err, it is evident that the emphasis here is more on “loosely” than on “captures”. At any rate, the craft of polling is a good case in point for the difficulty of practical random sampling.

Let us stay with the serious applications of random sampling for a short while before we turn to the frivolous ones. In the realm of scientific computing, the Monte Carlo method has a great demand for random samples, and this not only for numerical integration, but also for solving integral equations, boundary-value problems with partial differential equations, and linear-algebra problems involving matrices of large size, as well as for the optimization of functions and for many other tasks. Ever since the invention of the Monte Carlo method, problems of computational physics were treated by this method, a prominent example being particle transport through a solid medium in nuclear physics. This problem is of crucial importance for the safety of nuclear reactors.

Monte Carlo methods belong to the wider family of simulation methods which strive to gain information about complex and large-scale systems by random sampling. For instance, you may think of the management of production processes in a big factory or flight-scheduling problems for a global airline. In these examples, the supply of resources, the demand for the company’s products, or the preferences and frequency of passengers are random processes that have to be simulated.

There is also the area of probabilistic algorithms in scientific computing and computer science where random samples are required. A probabilistic algorithm is basically like any other computational procedure inasmuch as it follows well-defined steps that can be programmed in software, but in certain steps of the algorithm we are allowed to make random choices. These choices come from a specified set and are produced by random sampling. For some computational tasks, probabilistic algorithms tend to arrive at the desired answer faster than conventional deterministic algorithms. In this book, you can find examples of probabilistic algorithms in Subsections 2.3.3, 2.4.2, 6.5.1, 6.6.1, and 6.6.2.

But believe it or not, the biggest consumer of random samples nowadays is the gaming industry, by which we mean not only things like slot machines in casinos, but also computer games. Take the typical slot machine: it contains an electronic device that selects “at random” one out of the several thousand possible combinations of fruits and other objects on the display. As a matter of fact, somebody has programmed this device and it runs through an algorithm that produces the supposedly random outcomes in real time. If you can hack this algorithm, then you could make millions at your local casino, provided the management does not become suspicious of your lucky streak and bans you from the premises. Similar features appear in computer games where various scenarios are selected “randomly”, but in reality according to a deterministic algorithm unbeknown to the user.

With these numerous applications of random sampling, it is evident that much thought has been spent on the actual generation of random samples. At the beginning of this subsection, we described the framework for random sampling in an abstract manner, namely a set with at least two elements and a probability distribution on it. In practice, the set from which the random samples are drawn will be of a concrete nature, for instance, the set $\{0,1\}$ of bits, a finite set of integers, the set $\mathbb{R}$ of real numbers, or a set of points in a Euclidean space. The problem of sampling from such typical concrete sets can usually be reduced to that of sampling from $\mathbb{R}$. In the latter case, we speak of *random number generation.*

Using statistical terminology, the task of random number generation presents itself in the following form: given a target distribution function F on $\mathbb{R}$, generate a sequence of real numbers that simulates a sequence of independent and identically distributed random variables with distribution function F. You may think of a *distribution function* F as a real-valued nondecreasing function on $\mathbb{R}$ with $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$. The intuitive meaning of the distribution function F is that, for all $x \in \mathbb{R}$, the probability $\Pr(r \leq x)$ that a random number r satisfies $r \leq x$ is equal to $F(x)$. We do not specify the mathematical definition of “independent” and rather appeal again to intuition: independence signifies that the choice of a random sample or of a random number is not influenced by previous choices of random samples or of random numbers in the same sampling procedure. Just think of so-called fair coin tosses for an illustration: if you know the outcomes of a run of fair coin tosses, then this should not give you any clue about the outcome of the next fair coin toss.

It is customary to break up the task of random number generation into two steps: (i) generate random numbers for an easy standardized distribution function on $\mathbb{R}$; (ii) transform the random numbers in step (i) into random numbers with the given target distribution function F on $\mathbb{R}$. As the easy standardized distribution function we choose the *uniform distribution function* U on $\mathbb{R}$ defined by $U(x) = 0$ for $x < 0$, $U(x) = x$ for $0 \leq x \leq 1$, and $U(x) = 1$ for $x > 1$ (see Figure 5.1). Random numbers whose target distribution function is U are called *uniform random numbers.* For uniform random numbers r we have $\Pr(r < 0) \leq \Pr(r \leq 0) = U(0) = 0$ and $\Pr(r > 1) = 1 - \Pr(r \leq 1) = 1 - U(1) = 0$, and so we can assume that uniform random numbers belong to the interval $[0,1]$. We emphasize that uniform random numbers satisfy the property that, for all $0 < x \leq 1$, the probability of a uniform random number from $[0,1]$ falling into the subinterval $[0,x]$ is equal to x. There is of course a formal similarity here with the concept of a uniformly distributed sequence (compare with Theorem 4.1.6), and this explains why the theory of uniform distribution of sequences plays a role in the analysis of uniform (pseudo)random numbers as we shall see.

In this book, we focus on uniform (pseudo)random numbers since it is in this area where the applications of number theory occur. The transformation step (ii) listed above involves the theory of special functions and elementary statistics, but

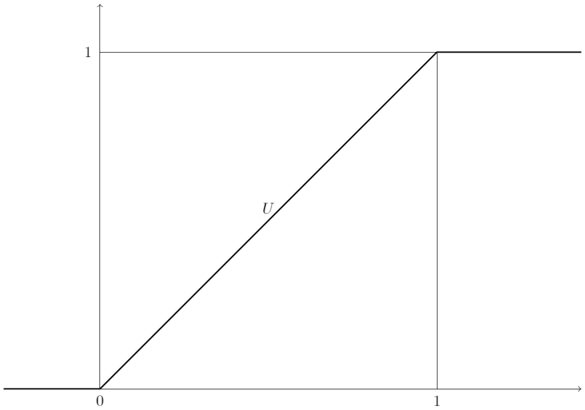

Figure 5.1: The graph of U

no number theory. Therefore we just say a few words about step (ii). Consider the case where the target distribution function F is strictly increasing and continuous on $\mathbb{R}$. Then it is clear that the image of F is the open interval $(0, 1)$ and that the inverse function $F^{-1} : (0, 1) \to \mathbb{R}$ exists. Since uniform random numbers attain the values 0 and 1 with probability 0, we can assume here that the uniform random numbers $r_1, r_2, \dots$ generated in step (i) above lie in the open interval $(0, 1)$. Now we determine a sequence $x_1, x_2, \dots$ of real numbers by $x_n = F^{-1}(r_n)$ for all $n \geq 1$. Then

$$\Pr(x_n \leq x) = \Pr(F^{-1}(r_n) \leq x) = \Pr(r_n \leq F(x)) = F(x)$$

for all $x \in \mathbb{R}$, and so $x_1, x_2, \dots$ can be viewed as a sequence of random numbers with target distribution function F. For obvious reasons, this transformation method is called the *inversion method*.

Example 5.1.1 Consider a so-called Cauchy distribution function F which is defined by

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan(x/\sigma) \qquad \text{for all } x \in \mathbb{R},$$

where σ is a positive constant. Then F is strictly increasing and continuous on $\mathbb{R}$, $\lim_{x \to -\infty} F(x) = 0$, and $\lim_{x \to \infty} F(x) = 1$. Hence F is a distribution function for which we can apply the inversion method. The inverse function F^{-1} of F is given by

$$F^{-1}(x) = \sigma \tan\left(\pi \left(x - \frac{1}{2}\right)\right) \qquad \text{for } 0 < x < 1.$$

In this case, the function values of F^{-1} can be efficiently computed by using standard mathematical software.

There are many situations where there is no nice closed-form expression for the inverse function F^{-1} of a given distribution function F, and in such cases the inversion method is not practical. However, various other methods for transforming uniform random numbers into random numbers with a prescribed nonuniform target distribution function are available. Methods geared to common distribution functions such as the family of normal distribution functions were extensively studied and can be very efficient. The voluminous book [36] is devoted entirely to these transformation methods, and a more recent treatment of this topic can be found in [71].

The first attempts to generate (uniform) random numbers on a large scale used physical devices such as automated roulette wheels, gadgets producing white noise, and counts of the emission of radiated particles. There was the legendary machine ERNIE (Electronic Random Number Indicator Equipment) that picked the winning numbers in the British Premium Bonds lottery for many years. Some machine-generated random numbers were even published in the form of tables, such as those in the RAND tables [164] that were widely employed in their time, but printing random numbers on paper and thus codifying them for eternity somehow seems to defeat the conventional idea of random numbers. An account of physical methods for the generation of random numbers is given in the book edited by Shreider [185, Chapter VI].

However, the utilization of physically generated random numbers is problematic. In the first place, the generated random numbers may have a small bias because of measurement errors due to imperfect equipment or because the underlying physical process is not sufficiently well understood and actually follows a probability distribution that is somewhat different from the anticipated one. Extensive statistical testing is mandatory in order to detect a possible bias (see Subsection 5.1.2 for such tests). Then there is the crucial issue of reproducibility: just like laboratory experiments, scientific computations must be reproducible by other experts for the purpose of verification. For computations involving physically generated random numbers, this means that the random numbers have to be stored so that they are available when the same computation is repeated. A high-level Monte Carlo computation with many runs may consume up to 10^{12} random numbers, and the storage of so many random numbers can be cumbersome and error-prone.

In view of the inconvenience of physically generated random numbers, practitioners switched to deterministic algorithms that generate random numbers in a quick and user-friendly way in a computer and depend only on a few input parameters. Even large bulks of random numbers can then be produced "on the fly" as one says, that is, in real time as they are needed, and so the problems of reproducibility and storage are vanquished. It is an additional advantage of such computer-generated random numbers that they can often be subjected to a rigorous theoretical analysis, and this will be illustrated in the present chapter. Suitable theoretical results may alleviate to some extent the need for time-consuming statistical testing of computer-generated random numbers. To be sure, it is of course an oxymoron to speak of *random* numbers generated by a *deterministic* algorithm. Therefore the terminology *pseudorandom numbers* is frequently used for computer-generated random numbers. We follow this terminology as it allows for a tidy differentiation from "true" random numbers (whatever these may be). Furthermore, we restrict the usage to *uniform* pseudorandom numbers. Thus, in this book, pseudorandom numbers are numbers that are generated by a deterministic algorithm and that attempt to simulate the uniform distribution function U on $\mathbb{R}$.

The deterministic algorithms for generating pseudorandom numbers, and often also certain parameters in these algorithms, have to be chosen very carefully in order to arrive at pseudorandom numbers of good quality. There is a long history of bad choices of algorithms and parameters in this area, and we will of course steer away from these known bad choices. A memorable lesson about the judicious choice of algorithms is contained in the title of the classical paper [29]: "Random number generation is too important to be left to chance". Or to quote the leading computer scientist Donald Knuth [81, Section 3.1]: "Random numbers should not be generated with a method chosen at random. Some theory should be used."

5.1.2 Testing pseudorandom numbers

Pseudorandom numbers are generated by deterministic algorithms, and from this angle there is *a priori* absolutely no guarantee that they will do what they are supposed to do, namely to simulate the uniform distribution function U on $\mathbb{R}$ and to possess desirable statistical independence properties. Therefore quality control is indispensable in the business of pseudorandom number generation. As the famous dictum ascribed alternately to Mark Twain or to Lenin or to Ronald Reagan says: "Confidence is good, but control is better."

There are several categories according to which we can assess the quality of pseudorandom numbers, such as statistical, structural, and complexity-theoretic criteria. One may also add ease of implementation and speed of the algorithm producing pseudorandom numbers, but in the age of high-speed computers the time spent on generating pseudorandom numbers is minuscule. Already in 1990 the computer expert Fred James at the nuclear research center CERN in Geneva noted in his paper [73]: "This [efficiency] was considered very important in the early days, but with the kind of computations being performed now, both the computer time and memory space taken by random number generation are increasingly insignificant and can almost always be neglected." We will discuss several statistical tests for pseudorandom numbers in this subsection. Structural criteria refer to aspects such as period length and lattice structure, and we will investigate such properties for specific methods of pseudorandom number generation. Complexity-theoretic requirements are not so important for pseudorandom numbers used in Monte Carlo methods and simulation methods, but they are essential in pseudorandom bit generation for cryptography (see Section 5.4).

The statistical testing of pseudorandom numbers operates in a setting where we are given a sequence $x_0, x_1, \ldots$ of pseudorandom numbers in the interval $[0, 1)$ and a large integer N. The statistical tests use the first N terms $x_0, x_1, \ldots, x_{N-1}$ of the sequence or a slightly longer initial segment of the sequence. In each test we compute a certain test quantity and compare it with a benchmark, namely the value of the test quantity for a "truly random" sequence. The benchmark value is usually obtained by probability theory.

An absolute must is the *uniformity test* (or *equidistribution test*) which checks whether the given pseudorandom numbers really follow our standardized target distribution function, namely the uniform distribution function U on $\mathbb{R}$. To this end, we calculate the star discrepancy D_N^* of the numbers $x_0, x_1, \ldots, x_{N-1}$. Note that the definition of the star discrepancy in Definition 4.1.11 indicates that D_N^* represents the maximal deviation between the actual distribution of the numbers $x_0, x_1, \ldots, x_{N-1}$ and the ideal distribution function U. There is a law of the iterated logarithm for the star discrepancy due to Chung [25] which says that

$$\limsup_{N \to \infty} \frac{(2N)^{1/2} D_N^*(\mathcal{S})}{(\log \log N)^{1/2}} = 1$$

with probability 1, that is, for a "truly random" sequence $\mathcal{S}$ of points in $[0, 1)$. In particular $D_N^*(\mathcal{S}) = O\left(N^{-1/2}(\log \log N)^{1/2}\right)$ for all $N \geq 3$, where the implied constant may depend on the sequence $\mathcal{S}$. For some algorithms for pseudorandom number generation, the star discrepancy D_N^* can be bounded by means of a mathematical theorem, and then no computations have to be performed for the uniformity test.

The *permutation test* examines the relative ordering among successive pseudorandom numbers. We choose an integer $s \geq 2$ and form the s-tuples

$$(x_n, x_{n+1}, \ldots, x_{n+s-1}) \quad \text{for } n = 0, 1, \ldots, N-1.$$

Note that here the first $N + s - 1$ terms of the given sequence of pseudorandom numbers are needed, and similar statements hold for the following statistical tests. There are $s!$ possible relative orderings among the entries of such an s-tuple and these orderings are equiprobable. We determine the frequency of each ordering, with some convention for breaking ties of entries, and we use the maximal deviation of these frequencies from the expected number of occurrences as the basis for a statistical test.

A very popular test, not only for pseudorandom numbers but also in other applications of statistics, is the *serial correlation test*. This is a test for the interdependence between x_n and x_{n+h}, where $h \geq 1$ is a given integer. The test is performed by calculating the *serial correlation coefficient*

$$\sigma_N^{(h)} := \frac{M_N(x_n x_{n+h}) - (M_N(x_n))^2}{M_N(x_n^2) - (M_N(x_n))^2},$$

where $M_N(v_n) = N^{-1}\sum_{n=0}^{N-1} v_n$ denotes the mean value of real numbers $v_0, v_1, \dots, v_{N-1}$ and where we assume that the denominator is nonzero. If x_n and x_{n+h} are statistically almost independent, then the absolute value $|\sigma_N^{(h)}|$ is small. It is a drawback of this test that the converse does not necessarily hold. The serial correlation test is widely used since serial correlation coefficients can be computed quickly.

The *serial test* is a more severe test for the statistical independence of successive pseudorandom numbers and is a multidimensional version of the uniformity test (see above). For a fixed dimension $s \geq 2$, we put

$$\mathbf{x}_n = (x_n, x_{n+1}, \dots, x_{n+s-1}) \in [0,1)^s \qquad \text{for } n = 0, 1, \dots, N-1.$$

Let $\mathcal{P}_N^{(s)}$ be the point set consisting of the points $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{N-1}$. Then we consider the star discrepancy $D_N^*(\mathcal{P}_N^{(s)})$ of this point set. The appropriate statistical benchmark result is the s-dimensional law of the iterated logarithm for the star discrepancy due to Kiefer [79]. This law affirms that

$$\limsup_{N\to\infty} \frac{(2N)^{1/2} D_N^*(\mathcal{S})}{(\log\log N)^{1/2}} = 1$$

with probability 1, that is, for a "truly random" sequence $\mathcal{S}$ of points in $[0,1)^s$. In particular $D_N^*(\mathcal{S}) = O\left(N^{-1/2}(\log\log N)^{1/2}\right)$ for all $N \geq 3$, where the implied constant may depend on the sequence $\mathcal{S}$. As for the uniformity test, we can give theoretical bounds on the star discrepancy $D_N^*(\mathcal{P}_N^{(s)})$ in some cases.

Remark 5.1.2 It is intuitively clear that the serial test is stronger than the serial correlation test, in the sense that a small star discrepancy implies a small serial correlation coefficient, and we can also support this by a formal argument. For simplicity, we consider only the serial correlation coefficient $\sigma_N^{(h)}$ with $h = 1$, but we can proceed similarly for any integer $h \geq 1$. Let $x_0, x_1, \dots, x_{N-1}$ be pseudorandom numbers in $[0,1)$ with small star discrepancy D_N^*, say $D_N^* \leq \frac{1}{100}$. First we inspect the denominator

$$\mathrm{Den}(\sigma_N^{(1)}) = M_N(x_n^2) - (M_N(x_n))^2$$

of $\sigma_N^{(1)}$. By simple algebraic manipulations, we get

$$\begin{aligned}\mathrm{Den}(\sigma_N^{(1)}) &= \frac{1}{N}\sum_{n=0}^{N-1} x_n^2 - \Big(\frac{1}{N}\sum_{n=0}^{N-1} x_n\Big)^2 \\ &= \frac{1}{12} + \Big(\frac{1}{N}\sum_{n=0}^{N-1} x_n^2 - \frac{1}{3}\Big) - \Big(\frac{1}{N}\sum_{n=0}^{N-1} x_n - \frac{1}{2}\Big) - \Big(\frac{1}{N}\sum_{n=0}^{N-1} x_n - \frac{1}{2}\Big)^2.\end{aligned}$$

The last three bracketed expressions can be bounded by the Koksma inequality (see Theorem 4.1.21). Note that an increasing function g on $[0,1]$ has bounded variation

$V(g) = g(1) - g(0)$, and so the functions $g_1(x) = x$ and $g_2(x) = x^2$ on $[0, 1]$ satisfy $V(g_1) = V(g_2) = 1$. Consequently,

$$\mathrm{Den}(\sigma_N^{(1)}) \geq \frac{1}{12} - 2D_N^* - (D_N^*)^2 \geq \frac{1}{12} - \frac{2}{100} - \left(\frac{1}{100}\right)^2 > \frac{1}{16}. \tag{5.1}$$

For the numerator $\mathrm{Num}(\sigma_N^{(1)})$ of $\sigma_N^{(1)}$ we obtain

$$\begin{aligned} |\mathrm{Num}(\sigma_N^{(1)})| &= |M_N(x_n x_{n+1}) - (M_N(x_n))^2| \\ &\leq \Big|M_N(x_n x_{n+1}) - \frac{1}{4}\Big| + \Big|(M_N(x_n))^2 - \frac{1}{4}\Big| \\ &\leq \Big|M_N(x_n x_{n+1}) - \frac{1}{4}\Big| + \frac{3}{2}\Big|M_N(x_n) - \frac{1}{2}\Big| \\ &\leq \Big|M_N(x_n x_{n+1}) - \frac{1}{4}\Big| + \frac{3}{2}D_N^*. \end{aligned}$$

We can bound the term $\left|M_N(x_n x_{n+1}) - \frac{1}{4}\right|$ by the Koksma-Hlawka inequality (see Theorem 4.1.41). We introduce the point set $\mathcal{P}_N^{(2)}$ consisting of the points

$$\mathbf{x}_n = (x_n, x_{n+1}) \in [0, 1)^2 \qquad \text{for } n = 0, 1, \ldots, N-1$$

and the function $f(x, y) = xy$ on $[0, 1]^2$. The variation $V(f)$ of f on $[0, 1]^2$ in the sense of Hardy and Krause is given by

$$V(f) = \int_0^1 \int_0^1 \Big|\frac{\partial^2 f(x, y)}{\partial x \partial y}\Big| dx dy + \int_0^1 \Big|\frac{df(x, 1)}{dx}\Big| dx + \int_0^1 \Big|\frac{df(1, y)}{dy}\Big| dy = 3$$

according to (4.25), and so the Koksma-Hlawka inequality produces the bound

$$\Big|M_N(x_n x_{n+1}) - \frac{1}{4}\Big| = \Big|\frac{1}{N}\sum_{n=0}^{N-1} f(\mathbf{x}_n) - \int_0^1 \int_0^1 f(x, y) dx dy\Big| \leq 3D_N^*(\mathcal{P}_N^{(2)}).$$

If we use also the inequality $D_N^* \leq D_N^*(\mathcal{P}_N^{(2)})$ which is derived from the projection principle in Remark 4.1.37, then we obtain

$$|\mathrm{Num}(\sigma_N^{(1)})| \leq 3D_N^*(\mathcal{P}_N^{(2)}) + \frac{3}{2}D_N^* \leq \frac{9}{2}D_N^*(\mathcal{P}_N^{(2)}).$$

Together with the lower bound on $\mathrm{Den}(\sigma_N^{(1)})$ in (5.1), this yields

$$|\sigma_N^{(1)}| = \frac{|\mathrm{Num}(\sigma_N^{(1)})|}{|\mathrm{Den}(\sigma_N^{(1)})|} < 16 \cdot \frac{9}{2}D_N^*(\mathcal{P}_N^{(2)}) = 72D_N^*(\mathcal{P}_N^{(2)}).$$

Hence a small star discrepancy $D_N^*(\mathcal{P}_N^{(2)})$ implies a small serial correlation coefficient $\sigma_N^{(1)}$.

There are many more statistical tests for pseudorandom numbers, and some experts have even designed entire batteries of tests such as the DIEHARD test battery (see [54, Section 6.2]). The name of this test battery is an amusing pun since DieHard is a popular brand of car batteries in the United States, but maybe it refers also to the eponymous series of action movies starring Bruce Willis in which his stamina is tested in the extreme. The classical treatment of statistical tests for pseudorandom numbers is given in the book of Knuth [81, Section 3.3].

It should be evident that a deterministic sequence of numbers cannot perform well under *all* conceivable tests for randomness. Therefore the user of pseudorandom numbers should be aware of the specific desirable statistical properties of the random numbers for the computational task at hand and should choose pseudorandom numbers that are known to pass the corresponding statistical tests. For instance, if all that is needed is the statistical independence of any two successive uniform random numbers, then pseudorandom numbers passing the two-dimensional serial test are quite sufficient for this particular purpose.

5.2 The linear congruential method

5.2.1 Basic properties

It is striking that practically all currently employed methods for generating uniform pseudorandom numbers use number-theoretic or algebraic techniques. One of the first methods in the history of pseudorandom number generation was the *linear congruential method* which was introduced by the number theorist Derrick H. Lehmer at a conference at Harvard University in 1949 (see [96]). This method is still popular because of its simplicity. We need only two parameters for this method, namely a large integer m and an integer a with $\gcd(a, m) = 1$. Let $Z_m = \{0, 1, \dots, m-1\}$ again denote the least residue system modulo m. Then we choose an initial value $z_0 \in Z_m$ with $\gcd(z_0, m) = 1$ and we generate a sequence $z_0, z_1, \dots$ of elements of Z_m by the recursion

$$z_{n+1} \equiv a z_n \pmod{m} \qquad \text{for } n = 0, 1, \dots. \tag{5.2}$$

From this sequence, we derive the *linear congruential pseudorandom numbers*

$$x_n = \frac{z_n}{m} \in [0, 1) \qquad \text{for } n = 0, 1, \dots. \tag{5.3}$$

In this context, m is referred to as the *modulus* and a as the *multiplier*. Since a matters only modulo m, we assume that $a \in Z_m$.

These definitions immediately yield some simple consequences. For instance, the recursion (5.2) for the integers z_n implies the explicit formula

$$z_n \equiv a^n z_0 \pmod{m} \qquad \text{for } n = 0, 1, \dots. \tag{5.4}$$

Then the pseudorandom numbers x_n are given by the fractional parts

$$x_n = \left\{ \frac{1}{m} a^n z_0 \right\} \qquad \text{for } n = 0, 1, \ldots . \tag{5.5}$$

Let T be the multiplicative order of a modulo m, that is, T is the least positive integer h such that $a^h \equiv 1 \pmod{m}$. Then $z_{n+T} = z_n$ for all $n \geq 0$, and so the sequence $(z_n)_{n=0}^{\infty}$ is purely periodic with period length T. Since $\gcd(z_0, m) = 1$, the number T is actually the least period length. The numbers x_n are obtained from the integers z_n by multiplying by the constant $1/m$, and so the sequence $(x_n)_{n=0}^{\infty}$ has the same least period length. In general, for a periodic sequence $(w_n)_{n=0}^{\infty}$ we write $\mathrm{per}(w_n)$ for the least period length of $(w_n)_{n=0}^{\infty}$. We summarize the information about the least period length of $(x_n)_{n=0}^{\infty}$ as follows.

Proposition 5.2.1 *Let $(x_n)_{n=0}^{\infty}$ be a sequence of linear congruential pseudorandom numbers with modulus m and multiplier a. Then $\mathrm{per}(x_n) = T$, where T is the multiplicative order of a modulo m.*

One may question whether it is prudent to admit pseudorandom numbers that are rational numbers with the same denominator m. Certainly, such numbers would be unlikely if we were to draw truly random samples from the uniform distribution on the interval $[0, 1]$. However, there is the pragmatic viewpoint that once we entrust a computer with the generation of pseudorandom numbers, then we have to live with the fact that such a machine can represent real numbers only with a finite precision. From this perspective, it is legitimate to work with rational pseudorandom numbers having a finite precision. For instance, if your computer has a 32-bit processor, then it is reasonable and practical to generate only rational numbers with denominator 2^{32}.

Another issue is that of the periodicity of sequences of linear congruential pseudorandom numbers. Again, truly random samples would definitely not exhibit periodic patterns, so there is a problem here from a scrupulous statistical point of view. But as before we adopt a pragmatic stance and we concede that if the period length is significantly larger than the number of pseudorandom numbers actually consumed in a computation, then the user will not "notice" the periodicity of the sequence of pseudorandom numbers. In the case of linear congruential pseudorandom numbers, we have $\mathrm{per}(x_n) = T \leq m$ on account of Proposition 5.2.1, and so the modulus m should be considerably larger than the total number of linear congruential pseudorandom numbers used in a computation. It must be added that every algorithm currently utilized in practice for pseudorandom number generation produces periodic sequences, and so this issue of periodicity is not something particular for the linear congruential method.

Now we address the question of the choice of the two parameters m and a in the linear congruential method. We have already observed that the modulus m

should be quite large, for reasons of a fine discretization of the interval $[0, 1]$ and of a sufficiently large period length. On the other hand, for ease of implementation it is convenient for the modulus to fit into the word size of the processor. For instance, for a 32-bit processor it is preferable to have $m \leq 2^{32}$. Furthermore, the theory of the linear congruential method is nicer if m is a prime number or a power of 2. Again for a 32-bit processor, popular choices for the modulus are the Mersenne prime $m = 2^{31} - 1$ (see Subsection 6.6.3 for Mersenne primes) or $m = 2^{32}$. For higher precision and period length, one may take $m = 2^{48}$ or $m = 2^{64}$ or the prime numbers $m = 2^{48} - 59$, $m = 2^{63} - 25$, or $m = 2^{64} - 59$.

The selection of good multipliers a is a subtler affair. The primary requirement is of course that the generated sequence of linear congruential pseudorandom numbers should have a large period length, and this means in view of Proposition 5.2.1 that the multiplicative order T of a modulo m should be large. Once m has been fixed, it is therefore reasonable to choose a such that T is as large as possible for this value of m. This is easy if m is a prime number, for then the largest possible value of T is $\phi(m) = m - 1$ and this value of T is attained if and only if a is a primitive root modulo m (see Definition 1.2.19). In the other interesting case, namely when m is a power of 2, the following result provides the desired information. Note that we can ignore very small values of m since they are irrelevant for our purpose.

Proposition 5.2.2 *Let $m = 2^k$ with an integer $k \geq 3$. Then the largest value of a multiplicative order modulo m is 2^{k-2}. For $k \geq 4$, an integer a has multiplicative order 2^{k-2} modulo m if and only if $a \equiv \pm 3 \pmod 8$.*

Proof. For every integer $k \geq 1$ and every odd integer a, let $t_k(a)$ be the multiplicative order of a modulo 2^k. Then $a^{t_k(a)} = 2^k b + 1$ for some $b \in \mathbb{Z}$, hence by squaring we get

$$a^{2t_k(a)} = 2^{2k}b^2 + 2^{k+1}b + 1 \equiv 1 \pmod{2^{k+1}},$$

and so

$$t_{k+1}(a) \leq 2t_k(a). \tag{5.6}$$

Since $t_3(a) \leq 2$ for all odd $a \in \mathbb{Z}$, it follows from (5.6) that $t_k(a) \leq 2^{k-2}$ for all $k \geq 3$. For odd $a \not\equiv 1 \pmod 8$ it is trivial that $t_3(a) = 2$, and so we can assume that $k \geq 4$ from now on. Every $a \equiv \pm 1 \pmod 8$ satisfies $t_4(a) \leq 2$, and so (5.6) implies in this case that $t_k(a) \leq 2^{k-3}$ for all $k \geq 4$. From $\phi(2^k) = 2^{k-1}$ we infer that $t_k(a)$ is always a power of 2 (compare with Remark 1.3.12). Thus, we can finish the proof by showing that $a^{2^{k-3}} \not\equiv 1 \pmod{2^k}$ for $a \equiv \pm 3 \pmod 8$. In fact, we prove for such a that

$$a^{2^{k-3}} \equiv 2^{k-1} + 1 \pmod{2^k} \qquad \text{for all } k \geq 4.$$

This is trivial for $k = 4$. If the assertion is shown for some $k \geq 4$, then $a^{2^{k-3}} = 2^k d + 2^{k-1} + 1 = 2^{k-1}(2d + 1) + 1$ for some $d \in \mathbb{Z}$, and so by squaring we obtain

$$a^{2^{k-2}} = 2^{2k-2}(2d + 1)^2 + 2^k(2d + 1) + 1 \equiv 2^k + 1 \pmod{2^{k+1}},$$

which completes the induction. □

Remark 5.2.3 In the literature one has also considered the so-called *inhomogeneous case* in the linear congruential method where the recursion (5.2) is replaced by

$$z_{n+1} \equiv az_n + c \pmod{m} \qquad \text{for } n = 0, 1, \ldots. \tag{5.7}$$

Here m is again a large integer and we choose $a, c \in Z_m$ with $\gcd(a, m) = 1$ and $c \neq 0$. The pseudorandom numbers x_n are again obtained by (5.3). This case adds of course a slight complication, but it is of interest since we can sometimes achieve $\mathrm{per}(x_n) = \mathrm{per}(z_n) = m$. For instance, let $m = 2^k$ with an integer $k \geq 3$, let $a \equiv 5 \pmod{8}$, and let c be odd. Then it follows from (5.7) by induction on n that

$$z_n \equiv a^n z_0 + \frac{a^n - 1}{a - 1} c \pmod{2^k} \qquad \text{for } n = 0, 1, \ldots.$$

This implies that

$$z_n - z_0 \equiv \frac{a^n - 1}{a - 1}((a - 1)z_0 + c) \pmod{2^k} \qquad \text{for } n = 0, 1, \ldots.$$

Since $(a-1)z_0 + c$ is odd, we have $z_n = z_0$ for some positive integer n if and only if 2^k divides $(a^n - 1)/(a - 1)$. Now $a - 1$ is divisible by 4 but not by 8, and so $z_n = z_0$ for some $n \in \mathbb{N}$ if and only if $a^n \equiv 1 \pmod{2^{k+2}}$. The least $n \in \mathbb{N}$ for which this holds is $n = 2^{(k+2)-2} = 2^k$ by Proposition 5.2.2, and so $\mathrm{per}(x_n) = \mathrm{per}(z_n) = 2^k = m$.

The performance of linear congruential pseudorandom numbers under the uniformity test in Subsection 5.1.2 is easy to describe if we apply this test to the full period and if the least period length is close to the modulus. Let $\mathcal{S} = (x_n)_{n=0}^{\infty}$ be a sequence of linear congruential pseudorandom numbers with modulus m and least period length $T = \mathrm{per}(x_n)$. Let $D_T^*(\mathcal{S})$ be the star discrepancy of the first T terms of $\mathcal{S}$, that is, $D_T^*(\mathcal{S})$ is the star discrepancy of the pseudorandom numbers in the full period. The simplest case is $T = m$, which can sometimes be achieved by using the inhomogeneous recursion (5.7) as shown in Remark 5.2.3. If $T = m$, then it is evident that the numbers $x_0, x_1, \ldots, x_{T-1}$ in the full period run exactly through all rational numbers in $[0, 1)$ with denominator m in some order. Therefore $D_T^*(\mathcal{S}) = 1/m$ by Proposition 4.1.16.

From now on, unless stated explicitly otherwise, we discuss only linear congruential pseudorandom numbers derived from the original homogeneous recursion (5.2). If m is a prime number and $T = m - 1$, then the numbers $x_0, x_1, \ldots, x_{T-1}$ are $1/m, 2/m, \ldots, (m-1)/m$ in some order. Therefore again $D_T^*(\mathcal{S}) = 1/m$ by Proposition 4.1.16. Another interesting case is $m = 2^k$ with an integer $k \geq 4$ and $a \equiv 5 \pmod{8}$. Then $T = 2^{k-2} = m/4$ by Proposition 5.2.2. Furthermore, (5.4) implies that $z_n \equiv z_0 \pmod{4}$ for all $n \geq 0$, and so the point set consisting of

$x_0, x_1, \dots, x_{T-1}$ is equal to the point set consisting of all rational numbers in $[0,1)$ of the form b/m with an integer $b \equiv z_0 \pmod 4$. Since $z_0 \equiv \pm 1 \pmod 4$, a straightforward computation based on Proposition 4.1.16 shows that $D_T^*(\mathcal{S}) = 3/m$.

If T is considerably smaller than the modulus m, then we cannot expect such simple explicit formulas for $D_T^*(\mathcal{S})$, but we can provide upper bounds on $D_T^*(\mathcal{S})$. We focus on the case where the modulus is a prime number, and we write as usual p for a prime number. Results for other moduli can also be obtained, but they are more complicated (see [125]). We need the following bound on exponential sums. For $M \in \mathbb{N}$ we put $\chi_M(z) = e^{2\pi \mathrm{i} z/M}$ for all $z \in \mathbb{Z}$.

Lemma 5.2.4 *Let p be a prime number and let $a, b, d \in \mathbb{Z}$ with* $\gcd(a,p) = \gcd(b,p) = 1$. *Let T be the multiplicative order of a modulo p. Then*

$$\Big|\sum_{n=0}^{T-1} \chi_p(ba^n)\chi_T(dn)\Big| \le \begin{cases} (p-T)^{1/2} & \text{if } d \equiv 0 \pmod T, \\ p^{1/2} & \text{otherwise.} \end{cases}$$

Proof. We use a method from the proof of Lemma 3.3.38. It is convenient to put

$$s(b,d) = \sum_{n=0}^{T-1} \chi_p(ba^n)\chi_T(dn).$$

The general term of this sum, viewed as a function of n, is periodic with period length T. Hence for every integer $r \ge 0$ we can write

$$s(b,d) = \sum_{n=0}^{T-1} \chi_p(ba^{n+r})\chi_T(d(n+r)),$$

and so

$$|s(b,d)| = \Big|\sum_{n=0}^{T-1} \chi_p(ba^r a^n)\chi_T(dn)\Big| = |s(ba^r, d)|.$$

Since the integers $b, ba, \dots, ba^{T-1}$ are pairwise incongruent modulo p and not divisible by p, it follows by putting $s(0,d) = \sum_{n=0}^{T-1} \chi_T(dn)$ that

$$\begin{aligned} T|s(b,d)|^2 &= \sum_{r=0}^{T-1} |s(ba^r,d)|^2 \\ &\le \sum_{g=1}^{p-1} |s(g,d)|^2 = \sum_{g=0}^{p-1} |s(g,d)|^2 - |s(0,d)|^2 \\ &= \sum_{h,j=0}^{T-1} \chi_T(d(h-j)) \sum_{g=0}^{p-1} \chi_p(g(a^h - a^j)) - |s(0,d)|^2 \\ &= pT - |s(0,d)|^2. \end{aligned}$$

Here in the penultimate step, we expanded $|s(g,d)|^2$ by using $|u|^2 = u\overline{u}$ for all $u \in \mathbb{C}$. Note also that in the last step, in the double sum over h and j only the terms with $h = j$ yield a nonzero contribution. Now $s(0,d) = T$ if $d \equiv 0 \pmod{T}$ and $s(0,d) = 0$ otherwise, and so we arrive at the desired bounds. □

Now we can establish an upper bound on the discrepancy $D_T(\mathcal{S})$, and therefore also on the star discrepancy $D_T^*(\mathcal{S})$, for a sequence $\mathcal{S}$ of linear congruential pseudorandom numbers with prime modulus.

Theorem 5.2.5 *Let $\mathcal{S} = (x_n)_{n=0}^{\infty}$ be a sequence of linear congruential pseudorandom numbers with prime modulus $p \geq 3$ and let $T = \mathrm{per}(x_n)$. Then*

$$D_T(\mathcal{S}) < \frac{(p-T)^{1/2}}{T}\Big(\log p + \frac{1}{3}\Big) + \frac{1}{p}.$$

Proof. We use Proposition 4.3.1 with $s = 1$ and $M = p$ together with the explicit formula (5.5) to obtain

$$D_T(\mathcal{S}) \leq \frac{1}{p} + \frac{1}{T} \sum_{h \in C^*(p)} \frac{1}{r(h,p)} \Big| \sum_{n=0}^{T-1} \chi_p(hz_0a^n) \Big|.$$

Now T is the multiplicative order of the multiplier a modulo p by Proposition 5.2.1, and so Lemma 5.2.4 with $d = 0$ yields

$$\Big| \sum_{n=0}^{T-1} \chi_p(hz_0a^n) \Big| \leq (p-T)^{1/2} \qquad \text{for all } h \in C^*(p).$$

We conclude that

$$D_T(\mathcal{S}) \leq \frac{1}{p} + \frac{(p-T)^{1/2}}{T} \sum_{h \in C^*(p)} \frac{1}{r(h,p)}.$$

We recall that $r(h,p) = p\sin(\pi|h|/p)$ for $h \in C^*(p)$, and so

$$\sum_{h \in C^*(p)} \frac{1}{r(h,p)} = \frac{2}{p} \sum_{h=1}^{(p-1)/2} \frac{1}{\sin(\pi h/p)} \leq \sum_{h=1}^{(p-1)/2} \frac{1}{h}$$

since $\sin(\pi u) \geq 2u$ for $0 \leq u \leq \frac{1}{2}$. Now

$$\sum_{h=1}^{(p-1)/2} \frac{1}{h} = 1 + \sum_{h=2}^{(p-1)/2} \frac{1}{h} \leq 1 + \int_1^{(p-1)/2} \frac{du}{u} = 1 + \log\frac{p-1}{2},$$

and therefore

$$\sum_{h \in C^*(p)} \frac{1}{r(h,p)} < \log p + \frac{1}{3}. \tag{5.8}$$

This completes the proof. □

We realize that the discrepancy bound in Theorem 5.2.5 does not yield an improvement on the trivial bound $D_T(\mathcal{S}) \leq 1$ if T is too small. In fact, T should be significantly larger than $p^{1/2}$ in order to obtain a nontrivial discrepancy bound. If T has the order of magnitude p, say $T = (p-1)/2$ or $T = (p-1)/4$, then the discrepancy bound in Theorem 5.2.5 is of the order of magnitude $p^{-1/2} \log p$, which is in good accordance with the law of the iterated logarithm for the uniformity test (see Subsection 5.1.2).

In applications of sequences of linear congruential pseudorandom numbers in Monte Carlo methods and simulation methods, we should use initial segments of the sequence that are shorter than the full period, since periodicity is excessively nonrandom and any influence of the periodicity property on the computation could prove ruinous. Therefore it is imperative that we study the discrepancies $D_N(\mathcal{S})$ and $D_N^*(\mathcal{S})$ also for N strictly less than the least period length T of the sequence $\mathcal{S}$ of linear congruential pseudorandom numbers. We now require a bound on exponential sums with $N < T$ terms.

Lemma 5.2.6 *Let $p \geq 3$ be a prime number and let $a, b \in \mathbb{Z}$ with $\gcd(a,p) = \gcd(b,p) = 1$. Let T be the multiplicative order of a modulo p and assume that $T \geq 2$. Then*

$$\Big| \sum_{n=0}^{N-1} \chi_p(ba^n) \Big| < p^{1/2} \Big(\log T + \frac{1}{3} \Big) + \frac{N}{T} (p-T)^{1/2} \quad \textit{for } 1 \leq N < T.$$

Proof. Our starting point is the identity

$$\sum_{n=0}^{N-1} \chi_p(ba^n) = \sum_{n=0}^{T-1} \chi_p(ba^n) \sum_{r=0}^{N-1} \frac{1}{T} \sum_{d=0}^{T-1} \chi_T(d(n-r)) \tag{5.9}$$

which holds since the innermost sum is equal to T if $n = r$ and equal to 0 if $n \neq r$. We rewrite this identity in the form

$$\sum_{n=0}^{N-1} \chi_p(ba^n) = \frac{1}{T} \sum_{d=0}^{T-1} \Big(\sum_{r=0}^{N-1} \chi_T(-dr) \Big) \Big(\sum_{n=0}^{T-1} \chi_p(ba^n) \chi_T(dn) \Big).$$

By taking absolute values, we get

$$\Big| \sum_{n=0}^{N-1} \chi_p(ba^n) \Big| \leq \frac{1}{T} \sum_{d=0}^{T-1} \Big| \sum_{r=0}^{N-1} \chi_T(dr) \Big| \Big| \sum_{n=0}^{T-1} \chi_p(ba^n) \chi_T(dn) \Big|,$$

and an application of Lemma 5.2.4 yields

$$\Big| \sum_{n=0}^{N-1} \chi_p(ba^n) \Big| \leq \frac{p^{1/2}}{T} \sum_{d=1}^{T-1} \Big| \sum_{r=0}^{N-1} \chi_T(dr) \Big| + \frac{N}{T} (p-T)^{1/2}. \tag{5.10}$$

Now for $1 \le d \le T-1$,

$$\Big|\sum_{r=0}^{N-1} \chi_T(dr)\Big| = \Big|\sum_{r=0}^{N-1} (e^{2\pi \mathrm{i} d/T})^r\Big| = \frac{|e^{2\pi \mathrm{i} dN/T}-1|}{|e^{2\pi \mathrm{i} d/T}-1|} \le \frac{1}{\sin(\pi d/T)},$$

and so

$$\sum_{d=1}^{T-1}\Big|\sum_{r=0}^{N-1} \chi_T(dr)\Big| \le \sum_{d=1}^{T-1} \frac{1}{\sin(\pi d/T)} \le 2\sum_{d=1}^{\lfloor T/2\rfloor} \frac{1}{\sin(\pi d/T)}.$$

Next we use $\sin(\pi u) \ge 2u$ for $0 \le u \le \frac{1}{2}$, and then as in the proof of Theorem 5.2.5 we obtain

$$\sum_{d=1}^{T-1}\Big|\sum_{r=0}^{N-1} \chi_T(dr)\Big| \le T\sum_{d=1}^{\lfloor T/2\rfloor} \frac{1}{d} < T(\log T + \frac{1}{3}).$$

By plugging this bound into (5.10), we arrive at the desired inequality. □

Theorem 5.2.7 *Let $\mathcal{S} = (x_n)_{n=0}^{\infty}$ be a sequence of linear congruential pseudorandom numbers with prime modulus $p \ge 3$ and let $T = \mathrm{per}(x_n) \ge 2$. Then for $1 \le N < T$,*

$$D_N(\mathcal{S}) < \frac{p^{1/2}}{N}\Big(\log p + \frac{1}{3}\Big)\Big(\log T + \frac{1}{3}\Big) + \frac{(p-T)^{1/2}}{T}\Big(\log p + \frac{1}{3}\Big) + \frac{1}{p}.$$

Proof. We proceed in analogy with the proof of Theorem 5.2.5. First of all, Proposition 4.3.1 now yields

$$D_N(\mathcal{S}) \le \frac{1}{p} + \frac{1}{N}\sum_{h \in C^*(p)} \frac{1}{r(h,p)}\Big|\sum_{n=0}^{N-1} \chi_p(hz_0a^n)\Big|.$$

Next Lemma 5.2.6 implies that

$$D_N(\mathcal{S}) < \frac{1}{p} + \frac{1}{N}\Big(p^{1/2}\Big(\log T + \frac{1}{3}\Big) + \frac{N}{T}(p-T)^{1/2}\Big)\sum_{h \in C^*(p)} \frac{1}{r(h,p)}.$$

Finally, an application of (5.8) produces the desired bound. □

Since always $T \le p-1$, the discrepancy bound in Theorem 5.2.7 can be written in the simplified form $D_N(\mathcal{S}) = O(N^{-1}p^{1/2}(\log p)^2)$. If T is equal to or close to $p-1$ and if N is of the same order of magnitude as T (for example if N is about one percent of T), then $D_N(\mathcal{S}) = O(p^{-1/2}(\log p)^2)$, which is in reasonably good accordance with the law of the iterated logarithm for the uniformity test (see Subsection 5.1.2).

5.2.2 Connections with good lattice points

The uniformity test for linear congruential pseudorandom numbers discussed in the preceding subsection is not a severe judge of multipliers. Note that the only way the multiplier enters into the formulas and bounds for the (star) discrepancy there is via the least period length T of the sequence of linear congruential pseudorandom numbers, or equivalently via the multiplicative order T of the multiplier a modulo m. However, for a fixed value of T there can be many multipliers with that multiplicative order modulo m, but the uniformity test does not discriminate between them. Even in the common case where the modulus is a prime number p and T has the largest possible value $T = p - 1$ for this modulus, very bad choices of multipliers a with $T = p - 1$ are possible (see Example 5.2.12 below).

We have to employ a more demanding statistical test in order to detect good multipliers and weed out bad ones, and such a test is the serial test (see Subsection 5.1.2). We focus on the case of a prime modulus p and a multiplier a that is a primitive root modulo p. Then the least period length T is equal to $p-1$ by Proposition 5.2.1. Let $\mathcal{S} = (x_n)_{n=0}^{\infty}$ be a corresponding sequence of linear congruential pseudorandom numbers. For a given integer $s \geq 2$, we analyze the s-dimensional serial test for these pseudorandom numbers. To this end, we form the points

$$\mathbf{x}_n = (x_n, x_{n+1}, \ldots, x_{n+s-1}) \in [0,1)^s \qquad \text{for } n = 0, 1, \ldots.$$

Since $x_n = z_n/p$ by (5.3), we can write $\mathbf{x}_n = (1/p)\mathbf{z}_n$ with

$$\mathbf{z}_n = (z_n, z_{n+1}, \ldots, z_{n+s-1}) \in Z_p^s \qquad \text{for } n = 0, 1, \ldots.$$

By using (5.4) and interpreting a congruence between vectors componentwise, we obtain

$$\mathbf{z}_n \equiv (a^n z_0, a^{n+1} z_0, \ldots, a^{n+s-1} z_0) \equiv a^n z_0 \mathbf{g}(a) \pmod{p} \quad \text{for } n = 0, 1, \ldots, \tag{5.11}$$

where

$$\mathbf{g}(a) = (1, a, a^2, \ldots, a^{s-1}) \in \mathbb{Z}^s.$$

Now we get a feeling of *déjà vu* because this lattice point is an old acquaintance: it is a lattice point of Korobov form introduced in (4.43) in Subsection 4.3.1. This observation is the beginning of an interesting story about the relationship between linear congruential pseudorandom numbers and good lattice points.

Let us follow this story further. First of all, we examine the full period of the given sequence $\mathcal{S} = (x_n)_{n=0}^{\infty}$ of linear congruential pseudorandom numbers, that is, we consider the first $p - 1$ terms of $\mathcal{S}$. Consequently, we take $n = 0, 1, \ldots, p - 2$ in (5.11). Now a is a primitive root modulo p and $\gcd(z_0, p) = 1$, and so for $n = 0, 1, \ldots, p - 2$ the coefficients $a^n z_0$ of $\mathbf{g}(a)$ on the right-hand side of (5.11) run modulo p through the set $\{1, \ldots, p - 1\}$ in some order. It follows that the point set consisting of $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{p-2}$ agrees with the point set comprising the fractional parts

$$\left\{\frac{n}{p}\,\mathbf{g}(a)\right\} \in [0,1)^s \qquad \text{for } n = 1, \ldots, p-1. \tag{5.12}$$

The point set (5.12), which we denote by $\mathcal{P}^*(\mathbf{g}(a), p)$, is nothing else but the point set $\mathcal{P}(\mathbf{g}, N)$ introduced in Subsection 4.3.1, with $\mathbf{g} = \mathbf{g}(a)$ and $N = p$, though with the origin deleted. Since $\mathcal{P}^*(\mathbf{g}(a), p)$ and $\mathcal{P}(\mathbf{g}(a), p)$ differ by only one point, it is fairly obvious that the discrepancies of these two point sets should be basically the same. The following lemma puts this in a quantitative form.

Lemma 5.2.8 *The point set $\mathcal{P}^*(\mathbf{g}(a), p)$ in* (5.12) *satisfies*

$$D_{p-1}(\mathcal{P}^*(\mathbf{g}(a), p)) \le \frac{p}{p-1} D_p(\mathcal{P}(\mathbf{g}(a), p)) + \frac{1}{p-1}.$$

Proof. Since $\mathcal{P}^*(\mathbf{g}(a), p)$ is $\mathcal{P}(\mathbf{g}(a), p)$ with the origin $\mathbf{0}$ deleted, we obtain

$$A(J; \mathcal{P}^*(\mathbf{g}(a), p)) = A(J; \mathcal{P}(\mathbf{g}(a), p)) - \varepsilon(J)$$

for every subinterval J of $[0,1)^s$, where $\varepsilon(J) = 1$ if $\mathbf{0} \in J$ and $\varepsilon(J) = 0$ if $\mathbf{0} \notin J$. Therefore

$$A(J; \mathcal{P}^*(\mathbf{g}(a), p)) - (p-1)\lambda_s(J) = (A(J; \mathcal{P}(\mathbf{g}(a), p)) - p\lambda_s(J)) + \lambda_s(J) - \varepsilon(J),$$

and so

$$|A(J; \mathcal{P}^*(\mathbf{g}(a), p)) - (p-1)\lambda_s(J)| \le p D_p(\mathcal{P}(\mathbf{g}(a), p)) + 1,$$

which yields the desired result. □

Consequently, we are in the agreeable situation that we can exploit all results on lattice point sets in Subsection 4.3.1. With $R(\mathbf{g}(a), p)$ given by Definition 4.3.2, we arrive for instance at the following discrepancy bound that pertains to the s-dimensional serial test for the full period of a sequence of linear congruential pseudorandom numbers with prime modulus p and with the multiplier a being a primitive root modulo p.

Theorem 5.2.9 *Let p be a prime number and let a be a primitive root modulo p. Then for every dimension $s \ge 2$, the discrepancy of the point set $\mathcal{P}^*(\mathbf{g}(a), p)$ in* (5.12) *satisfies*

$$D_{p-1}(\mathcal{P}^*(\mathbf{g}(a), p)) \le \frac{s+1}{p-1} + \frac{p}{2p-2} R(\mathbf{g}(a), p).$$

Proof. This follows from Theorem 4.3.3 and Lemma 5.2.8. □

Remember that we apply the s-dimensional serial test to our given sequence of linear congruential pseudorandom numbers because we want to discriminate between the various multipliers a that are primitive roots modulo p. The quantity $R(\mathbf{g}(a), p)$ is such a discriminator and it is small only for good choices of a. We could be tempted to apply Theorem 4.3.14 which implies that for every dimension $s \geq 2$ and every prime number p there exists an integer $a \in Z_p = \{0, 1, \ldots, p-1\}$ such that $R(\mathbf{g}(a), p) = O(p^{-1}(\log p)^s)$. But unfortunately there is no guarantee that a is a primitive root modulo p. Thus, we have to prove a version of Theorem 4.3.14 where a is confined to be a primitive root modulo p. We recall from Remark 1.4.35 that there are exactly $\phi(p-1)$ primitive roots modulo p in the least residue system Z_p modulo p.

Theorem 5.2.10 *Let $s \geq 2$ be a dimension and let p be a prime number. Then*

$$A_s(p) := \frac{1}{\phi(p-1)} \sum_{a \in Q(p)} R(\mathbf{g}(a), p) < \frac{s-1}{\phi(p-1)} (2 \log p + 2)^s,$$

where $Q(p)$ is the set of primitive roots modulo p in the least residue system modulo p.

Proof. Trivial modifications of the proof of Theorem 4.3.14 yield the result. □

Corollary 5.2.11 *For every dimension $s \geq 2$ and every prime number p, there exists a primitive root a modulo p such that the discrepancy of the point set $\mathcal{P}^*(\mathbf{g}(a), p)$ in* (5.12) *satisfies*

$$D_{p-1}(\mathcal{P}^*(\mathbf{g}(a), p)) < \frac{s+1}{p-1} + \frac{(s-1)p}{(2p-2)\phi(p-1)} (2 \log p + 2)^s.$$

Proof. This follows from Theorems 5.2.9 and 5.2.10. □

Since for every prime number $p \geq 3$ the bound $\phi(p-1) \geq cp/(\log \log p)$ holds with an absolute constant $c > 0$ (see [62, Chapter 18]), Corollary 5.2.11 demonstrates the existence of a primitive root a modulo p for which the discrepancy of $\mathcal{P}^*(\mathbf{g}(a), p)$ is at most of the order of magnitude $p^{-1}(\log p)^s \log \log p$. In fact, since this discrepancy bound is derived from an upper bound on the average value $A_s(p)$ in Theorem 5.2.10, it can be expected that a good proportion of the primitive roots a modulo p will lead to this order of magnitude for the discrepancy bound.

Example 5.2.12 There are many prime numbers p for which the integer 2 is a primitive root modulo p, for instance $p = 3, 5, 11, 13, 19, 29, 37, 53$, and so on. It follows in fact from a result of Hooley [68], which was proved under the assumption of the extended Riemann hypothesis, that for a positive proportion of all prime numbers p the integer 2 is a primitive root modulo p. Let us now take a prime number p for which 2 is a primitive root modulo p. Then we consider a sequence of linear congruential pseudorandom numbers with modulus p and multiplier $a = 2$. The least period length T of this sequence has the largest possible value $T = p-1$ for the modulus p. Now we apply the two-dimensional serial test to the full period of this sequence. This means that we study the discrepancy of the point set $\mathcal{P} = \mathcal{P}^*(\mathbf{g}(a), p)$ in (5.12) with $s = 2$ and $\mathbf{g}(a) = \mathbf{g}(2) = (1, 2) \in \mathbb{Z}^2$. Thus, $\mathcal{P}$ consists of the points $(n/p, \{2n/p\}) \in [0, 1)^2$ with $n = 1, \ldots, p-1$. We observe that the interval $J = [\frac{1}{4}, \frac{1}{2}) \times [0, \frac{1}{2})$ does not contain any point of $\mathcal{P}$, for if $\frac{n}{p} \in [\frac{1}{4}, \frac{1}{2})$, then $\frac{2n}{p} \in [\frac{1}{2}, 1)$. It follows that

$$D_{p-1}(\mathcal{P}) \geq \left| \frac{A(J;\mathcal{P})}{p-1} - \lambda_2(J) \right| = \lambda_2(J) = \frac{1}{8}$$

for all possible values of p. Hence the behavior of these linear congruential pseudorandom numbers under the two-dimensional serial test, and *a fortiori* under every higher-dimensional serial test, is really awful, whereas the behavior under the uniformity test is perfectly satisfactory. These pseudorandom numbers also yield catastrophic results in certain Monte Carlo computations. As an example, take the simple function $f(u_1, u_2) = \cos 2\pi(2u_1 - u_2)$ for $(u_1, u_2) \in [0, 1]^2$. The integral of f over $[0, 1]^2$ is equal to 0. On the other hand, $f(\mathbf{x}) = 1$ for all points $\mathbf{x}$ of $\mathcal{P}$. Thus, if we use as pseudorandom samples some points from $\mathcal{P}$, then the sample average is always equal to 1, no matter how many points we take. This very bad behavior is the fault of the multiplier since Corollary 5.2.11 shows that for the prime moduli considered in this example, there certainly exist very good choices for the multiplier. A rule of thumb can be deduced from this example, namely that a good multiplier should not be too small compared to the modulus.

Now we return to the point set $\mathcal{P}^*(\mathbf{g}(a), p)$ in (5.12). By an argument in the beginning of Subsection 4.3.2, all points of $\mathcal{P}^*(\mathbf{g}(a), p)$ lie on the s-dimensional lattice

$$L_s(a, p) = \bigcup_{n=1}^{p} \Big(\frac{n}{p}\, \mathbf{g}(a) + \mathbb{Z}^s \Big). \tag{5.13}$$

Thus, the points of $\mathcal{P}^*(\mathbf{g}(a), p)$ form a very regular pattern in the sense that they all fall on the lattice $L_s(a, p)$ (see Figure 5.2 for an illustration). In everyday language you could call this the structure of a grid. Marsaglia [110] expressed this memorably by the phrase "Random numbers fall mainly in the planes", a clever pun on a popular song from the musical *My Fair Lady*. Obviously, this lattice structure or grid structure is not at all what one would expect from truly random numbers and points. On account of this phenomenon, some practitioners are shying away from using linear congruential pseudorandom numbers in really challenging simulation problems.

The lattice $L_s(a, p)$ in (5.13), like any s-dimensional lattice, can be represented in the form (4.45) with s linearly independent vectors $\mathbf{b}_1, \ldots, \mathbf{b}_s \in \mathbb{R}^s$. In fact, for the

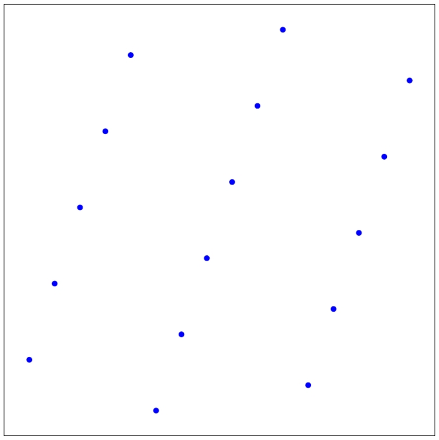

Figure 5.2: The point set $\mathcal{P}^*(\mathbf{g}(a), p)$ with $\mathbf{g}(a) = (1, 3)$ and $p = 17$

specific lattice $L_s(a, p)$ we can take $\mathbf{b}_1 = (1/p)\mathbf{g}(a)$ and $\mathbf{b}_i = \mathbf{e}_i$ for $2 \leq i \leq s$, where $\mathbf{e}_i$ is the ith vector in the standard ordered basis of $\mathbb{R}^s$, that is, $\mathbf{e}_2 = (0, 1, 0, \ldots, 0) \in \mathbb{R}^s, \ldots, \mathbf{e}_s = (0, \ldots, 0, 1) \in \mathbb{R}^s$. This information is very helpful when plotting, and also when analyzing, the lattice $L_s(a, p)$. Some researchers investigated the lattices of the form $L_s(a, p)$ and the point sets $\mathcal{P}^*(\mathbf{g}(a), p)$ from a geometric viewpoint and made recommendations of good multipliers a for the modulus p on this basis. Typical geometric criteria are the following: (i) the minimum number of parallel hyperplanes on which all points of $\mathcal{P}^*(\mathbf{g}(a), p)$ lie (this number should be as large as possible); (ii) the maximum distance between adjacent hyperplanes taken over all families of parallel hyperplanes that contain all points of $\mathcal{P}^*(\mathbf{g}(a), p)$ (this distance should be as small as possible). An easily readable account of this approach is given in [81, Section 3.3.4].

We emphasized in the context of the uniformity test for a sequence of linear congruential pseudorandom numbers that the discrepancy has to be investigated also for parts of the period of the sequence. The same holds of course for the serial test. The appropriate discrepancy bound can be established also for the case where the least period length T is less than $p - 1$, and the full period can be included in that case as well.

Theorem 5.2.13 *Let $(x_n)_{n=0}^{\infty}$ be a sequence of linear congruential pseudorandom numbers with prime modulus $p \geq 3$ and let $T = \mathrm{per}(x_n) \geq 2$. For a given dimension $s \geq 2$ and for an integer N with $1 \leq N \leq T$, let $\mathcal{P}$ be the point set consisting of the points*

$$\mathbf{x}_n = (x_n, x_{n+1}, \ldots, x_{n+s-1}) \in [0, 1)^s \qquad \textit{for } n = 0, 1, \ldots, N - 1.$$

Then

$$D_N(\mathcal{P}) < \frac{p^{1/2}}{N}\Big(\log p + \frac{4}{3}\Big)^s \Big(\log T + \frac{4}{3}\Big) + \frac{1}{2}R(\mathbf{g}(a), p) + \frac{s}{p}.$$

Proof. We start from the explicit formula

$$\mathbf{x}_n = \left\{ \frac{1}{p}\, a^n z_0 \mathbf{g}(a) \right\} \qquad \text{for } n = 0, 1, \ldots, N-1;$$

compare with (5.11). Then we apply Proposition 4.3.1 with $M = p$ and we obtain

$$D_N(\mathcal{P}) \le \frac{s}{p} + \frac{1}{N} \sum_{\mathbf{h} \in C_s^*(p)} \frac{1}{r(\mathbf{h}, p)} \Big| \sum_{n=0}^{N-1} \chi_p(a^n z_0 \mathbf{h} \cdot \mathbf{g}(a)) \Big|.$$

If $\mathbf{h} \cdot \mathbf{g}(a) \equiv 0 \pmod{p}$, then the last exponential sum is equal to N. Otherwise, we can use Lemmas 5.2.4 and 5.2.6 to get

$$\Big| \sum_{n=0}^{N-1} \chi_p(a^n z_0 \mathbf{h} \cdot \mathbf{g}(a)) \Big| < p^{1/2} \Big(\log T + \frac{1}{3} \Big) + \frac{N}{T} (p - T)^{1/2} < p^{1/2} \Big(\log T + \frac{4}{3} \Big).$$

Therefore

$$D_N(\mathcal{P}) < \frac{s}{p} + \frac{p^{1/2}}{N} \Big(\log T + \frac{4}{3} \Big) \sum_{\substack{\mathbf{h} \in C_s^*(p) \\ \mathbf{h} \cdot \mathbf{g}(a) \not\equiv 0 \ (\mathrm{mod}\ p)}} \frac{1}{r(\mathbf{h}, p)} + \sum_{\substack{\mathbf{h} \in C_s^*(p) \\ \mathbf{h} \cdot \mathbf{g}(a) \equiv 0 \ (\mathrm{mod}\ p)}} \frac{1}{r(\mathbf{h}, p)}.$$

Furthermore by (5.8),

$$\sum_{\substack{\mathbf{h} \in C_s^*(p) \\ \mathbf{h} \cdot \mathbf{g}(a) \not\equiv 0 \ (\mathrm{mod}\ p)}} \frac{1}{r(\mathbf{h}, p)} < \sum_{\mathbf{h} \in C_s(p)} \frac{1}{r(\mathbf{h}, p)} = \Big(1 + \sum_{h \in C^*(p)} \frac{1}{r(h, p)} \Big)^s < \Big(\log p + \frac{4}{3} \Big)^s.$$

Finally, as in the proof of Theorem 4.3.3 we obtain

$$\sum_{\substack{\mathbf{h} \in C_s^*(p) \\ \mathbf{h} \cdot \mathbf{g}(a) \equiv 0 \ (\mathrm{mod}\ p)}} \frac{1}{r(\mathbf{h}, p)} \le \frac{1}{2} \sum_{\substack{\mathbf{h} \in C_s^*(p) \\ \mathbf{h} \cdot \mathbf{g}(a) \equiv 0 \ (\mathrm{mod}\ p)}} \frac{1}{r(\mathbf{h})} = \frac{1}{2} R(\mathbf{g}(a), p),$$

and this completes the proof. □

We conclude from Theorem 5.2.13 that the multiplier a for the modulus p should be chosen in such a way that the quantity $R(\mathbf{g}(a), p)$ is small. An analogous rule holds for prime-power moduli. We refer to [128] for an in-depth discussion of the serial test for linear congruential pseudorandom numbers.

Example 5.2.14 Modern mathematical software typically advocates and employs good parameters in the linear congruential method. For instance, the GNU Scientific Library recommends among others the CRAY-system pseudorandom number generator RANF which uses the linear congruential method with modulus $m = 2^{48}$ and multiplier

$$a = 44485709377909.$$

This multiplier satisfies $a \equiv 5 \pmod{8}$, and so we get least period length 2^{46} by Propositions 5.2.1 and 5.2.2. If we use this multiplier in the inhomogeneous recursion (5.7) with an odd integer c, then we can achieve least period length 2^{48} according to Remark 5.2.3. This least period length exceeds by far the total number of pseudorandom numbers utilized in a routine simulation problem.

5.3 Nonlinear methods

5.3.1 The general nonlinear method

The lattice structure produced by linear congruential pseudorandom numbers (see Subsection 5.2.2) can be perceived as a deficiency of these pseudorandom numbers. This shortcoming becomes particularly pronounced if we make a bad choice of the multiplier, as we have seen with dramatic effect in Example 5.2.12. But even if we select the multiplier with care, the lattice structure is still there and can cause problems in Monte Carlo computations. For instance, consider the following generalization of Example 5.2.12.

Example 5.3.1 Let $\mathcal{S} = (x_n)_{n=0}^{\infty}$ be a sequence of linear congruential pseudorandom numbers with a prime modulus p and an arbitrary multiplier a satisfying $\gcd(a, p) = 1$. For an arbitrary dimension $s \geq 2$, let $\mathbf{h} = (h_1, \ldots, h_s) \in \mathbb{Z}^s$ be such that $\mathbf{h} \neq \mathbf{0}$ and $\mathbf{h} \cdot \mathbf{g}(a) \equiv 0 \pmod{p}$. With $\mathbf{u} = (u_1, \ldots, u_s)$, we introduce the function F on $[0, 1]^s$ given by

$$F(\mathbf{u}) = \cos 2\pi(h_1 u_1 + \cdots + h_s u_s) \qquad \text{for all } \mathbf{u} \in [0, 1]^s.$$

Then the integral of F over $[0, 1]^s$ is equal to 0. On the other hand, at all points

$$\mathbf{x}_n = (x_n, x_{n+1}, \ldots, x_{n+s-1}), \quad n = 0, 1, \ldots,$$

we get the function value $F(\mathbf{x}_n) = 1$ by (5.11), and so every sample average with these pseudorandom points has the value 1.

Further flaws of linear congruential pseudorandom numbers were pointed out in the literature. A case in point is the paper [43] where plausible geometric measures for the distribution of pseudorandom points in the unit square are studied and where it is revealed that pseudorandom points obtained from linear congruential pseudorandom numbers have a completely skew distribution with regard to these geometric measures. The book of Ripley [170, Sections 3.1 and 3.2] exposes some strange phenomena that arise when we use linear congruential pseudorandom numbers as inputs in certain algorithms for transforming uniform pseudorandom numbers into nonuniform pseudorandom numbers.

The problems with the linear congruential method stem from the simple linear nature of the recursion (5.2) at the heart of the method. In order to eliminate defects like the lattice structure, some features of nonlinearity should be introduced in the algorithms generating pseudorandom numbers. So why not replace the linear function of z_n on the right-hand side of (5.2) by a nonlinear function? This is the starting point of nonlinear methods for pseudorandom number generation. You may have heard of fractal geometry (its stars are the Mandelbrot set and Julia sets) which builds on a similar philosophy: iterating linear maps is boring, but iterating nonlinear maps can be exciting.

Let us now get down to business and implement this idea of a nonlinear method for generating pseudorandom numbers. For simplicity we choose a prime modulus p (which should again be large), and then we generalize (5.2) and generate elements $z_0, z_1, \dots$ of Z_p by choosing an initial value $z_0 \in Z_p$ and using the recursion

$$z_{n+1} \equiv \psi(z_n) \pmod{p} \qquad \text{for } n = 0, 1, \dots$$

with a map $\psi : Z_p \to \mathbb{Z}$. Since the values of ψ matter only modulo p, it suffices to view ψ as a map $\psi : Z_p \to Z_p$, that is, ψ is a self-map of Z_p. The recursion can then be written in the simpler form $z_{n+1} = \psi(z_n)$ for $n = 0, 1, \dots$. How do we figure out to what extent ψ is nonlinear? Here it is convenient to view the least residue system Z_p modulo p as the finite field $\mathbb{F}_p$ with p elements. Then there is the following nice description of self-maps of $\mathbb{F}_p$ which works in fact for *any* finite field $\mathbb{F}_q$.

Proposition 5.3.2 *Let q be a prime power and let $\psi : \mathbb{F}_q \to \mathbb{F}_q$ be a self-map of the finite field $\mathbb{F}_q$. Then there exists a uniquely determined polynomial $f \in \mathbb{F}_q[x]$ with* $\deg(f) < q$ *such that $\psi(c) = f(c)$ for all $c \in \mathbb{F}_q$.*

Proof. We consider the polynomial

$$f(x) = \sum_{b \in \mathbb{F}_q} \psi(b) \left(1 - (x-b)^{q-1}\right) \in \mathbb{F}_q[x].$$

Since $1 - a^{q-1}$ has the value 1 for $a = 0 \in \mathbb{F}_q$ and the value 0 for $a \in \mathbb{F}_q^*$, we get $\psi(c) = f(c)$ for all $c \in \mathbb{F}_q$. It is obvious that $\deg(f) < q$. If $g \in \mathbb{F}_q[x]$ is another polynomial with $\deg(g) < q$ such that $\psi(c) = g(c)$ for all $c \in \mathbb{F}_q$, then $(f-g)(c) = 0$ for all $c \in \mathbb{F}_q$. Hence the polynomial $f - g$ has at least q distinct roots. But $\deg(f-g) < q$, thus $f - g$ must be the zero polynomial, and so $f = g$. □

We are now ready to describe the final form of the *nonlinear* (*congruential*) *method* for the generation of pseudorandom numbers. Let p be a large prime number and select a polynomial $f \in \mathbb{F}_p[x]$. Then we generate a sequence $(z_n)_{n=0}^{\infty}$ of elements of $\mathbb{F}_p$ by choosing an initial value $z_0 \in \mathbb{F}_p$ and using the recursion

$$z_{n+1} = f(z_n) \qquad \text{for } n = 0, 1, \dots. \tag{5.14}$$

Since we do not want a linear method, we assume that $2 \le \deg(f) < p$. Now we identify $\mathbb{F}_p$ with Z_p and we derive *nonlinear* (*congruential*) *pseudorandom numbers* by setting

$$x_n = \frac{z_n}{p} \in [0, 1) \qquad \text{for } n = 0, 1, \dots. \tag{5.15}$$

As for the linear congruential method, the first issue is again the least period length. Since $\operatorname{per}(x_n) = \operatorname{per}(z_n)$, it suffices to study the periodicity properties of the sequence $(z_n)_{n=0}^{\infty}$. Note that in a full period of $(z_n)_{n=0}^{\infty}$ all terms are distinct because of the recursion (5.14), and so always $\mathrm{per}(z_n) \le p$. There is no simple criterion in terms of f for getting $\mathrm{per}(z_n) = p$. Contrary to the situation in the linear congruential method, the sequence $(z_n)_{n=0}^{\infty}$ may now have a preperiod.

Example 5.3.3 We can achieve $\mathrm{per}(z_n) = p$ by cheating a little bit. We put the cart before the horse, in the sense that we first construct a sequence $(z_n)_{n=0}^{\infty}$ and determine the polynomial $f \in \mathbb{F}_p[x]$ afterwards. Let $z_0, z_1, \ldots, z_{p-1}$ be a list of all elements of $\mathbb{F}_p$. By periodic continuation with period p we get the sequence $(z_n)_{n=0}^{\infty}$. In this way we guarantee that $\mathrm{per}(z_n) = p$. Now let ψ be the self-map of $\mathbb{F}_p$ defined by $\psi(z_n) = z_{n+1}$ for $0 \le n \le p-1$. By Proposition 5.3.2, ψ can be represented by a polynomial $f \in \mathbb{F}_p[x]$ with $\deg(f) < p$. It is clear from this construction that with this polynomial f, the recursion (5.14) generates the sequence $(z_n)_{n=0}^{\infty}$. Obviously, this example is only of academic interest, but at least it demonstrates that the value $\mathrm{per}(z_n) = p$ is attained for every prime number p with a suitable choice of f.

It is trivial that $\mathrm{per}(z_n) = p$ if and only if the full period of $(z_n)_{n=0}^{\infty}$ contains all elements of $\mathbb{F}_p$. In view of (5.14), this yields the simple necessary condition that the polynomial f must attain all values in $\mathbb{F}_p$. In other words, the self-map $c \in \mathbb{F}_p \mapsto f(c) \in \mathbb{F}_p$ of $\mathbb{F}_p$ must be surjective. Since this is a self-map of a finite set, it is surjective if and only if it is injective, and consequently it is surjective if and only if it is bijective. When it is injective, then it follows from (5.14) that the sequence $(z_n)_{n=0}^{\infty}$ is purely periodic, that is, there is no preperiod. Such bijective self-maps of finite fields are interesting in several applications, and so we introduce the following concept for arbitrary finite fields.

Definition 5.3.4 Let q be a prime power. A polynomial $f \in \mathbb{F}_q[x]$ for which the map $c \in \mathbb{F}_q \mapsto f(c) \in \mathbb{F}_q$ is bijective is called a *permutation polynomial* of $\mathbb{F}_q$.

Example 5.3.5 It is obvious that every linear polynomial over $\mathbb{F}_q$ is a permutation polynomial of $\mathbb{F}_q$. A power $x^k \in \mathbb{F}_q[x]$ with $k \ge 1$ is a permutation polynomial of $\mathbb{F}_q$ if and only if it maps a primitive element b of $\mathbb{F}_q$ into another primitive element of $\mathbb{F}_q$. Now b^k is a primitive element of $\mathbb{F}_q$ if and only if $\gcd(k, q-1) = 1$, and so x^k is a permutation polynomial of $\mathbb{F}_q$ if and only if $\gcd(k, q-1) = 1$. Since compositions of permutation polynomials of $\mathbb{F}_q$ are again permutation polynomials of $\mathbb{F}_q$, any polynomial $ax^k + c$ with $a \in \mathbb{F}_q^*$, $c \in \mathbb{F}_q$, $k \ge 1$, and $\gcd(k, q-1) = 1$ is a permutation polynomial of $\mathbb{F}_q$.

Remark 5.3.6 We recall that it is a necessary condition for $\mathrm{per}(z_n) = p$ that the polynomial f in (5.14) is a permutation polynomial of $\mathbb{F}_p$. However, this is not a sufficient condition. For a prime number $p \ge 5$, consider the polynomial $f(x) = x^{p-2} \in \mathbb{F}_p[x]$. Then f is a permutation polynomial of $\mathbb{F}_p$ by Example 5.3.5. The map $\psi : c \in \mathbb{F}_p \mapsto f(c) \in \mathbb{F}_p$ representing f satisfies $\psi(0) = 0$ and $\psi(c) = c^{-1}$ for $c \in \mathbb{F}_p^*$. Therefore $z_{n+2} = \psi(\psi(z_n)) = z_n$ for all $n \ge 0$, and so $\mathrm{per}(z_n) \le 2$.

It is, on first glance, somewhat surprising that certain degrees are excluded from the degrees of permutation polynomials of a given finite field.

Proposition 5.3.7 *Let q be a prime power. Then a polynomial $f \in \mathbb{F}_q[x]$ with $\deg(f) = d \geq 2$ and d dividing $q-1$ cannot be a permutation polynomial of $\mathbb{F}_q$.*

Proof. First we show that

$$\sum_{c \in \mathbb{F}_q} c^k = 0 \qquad \text{for } k = 0, 1, \ldots, q-2. \tag{5.16}$$

This is trivial for $k = 0$ with the standard convention $0^0 = 1 \in \mathbb{F}_q$. For $1 \leq k \leq q-2$ we choose a primitive element b of $\mathbb{F}_q$, and using $b^k \neq 1 \in \mathbb{F}_q$ for $1 \leq k \leq q-2$ we obtain

$$\sum_{c \in \mathbb{F}_q} c^k = \sum_{c \in \mathbb{F}_q^*} c^k = \sum_{j=0}^{q-2} (b^j)^k = \sum_{j=0}^{q-2} (b^k)^j = \frac{b^{k(q-1)} - 1}{b^k - 1} = 0.$$

Now we suppose that $f \in \mathbb{F}_q[x]$ with $\deg(f) = d \geq 2$ and d dividing $q-1$ were a permutation polynomial of $\mathbb{F}_q$. Then

$$\sum_{c \in \mathbb{F}_q} f(c)^{(q-1)/d} = \sum_{c \in \mathbb{F}_q} c^{(q-1)/d} = 0$$

by (5.16). We write $f(x)^{(q-1)/d} = ax^{q-1} + g(x)$ with $a \in \mathbb{F}_q^*$, $g \in \mathbb{F}_q[x]$, and $\deg(g) \leq q-2$. Then again by (5.16),

$$\sum_{c \in \mathbb{F}_q} f(c)^{(q-1)/d} = \sum_{c \in \mathbb{F}_q} \left(ac^{q-1} + g(c)\right) = a \sum_{c \in \mathbb{F}_q} c^{q-1}.$$

Finally,

$$\sum_{c \in \mathbb{F}_q} c^{q-1} = \sum_{c \in \mathbb{F}_q^*} c^{q-1} = \sum_{c \in \mathbb{F}_q^*} 1 = -1,$$

and we arrive at a contradiction. □

Now we return to the recursion (5.14) with $2 \leq d := \deg(f) < p$ and we note again that if $\mathrm{per}(z_n) = p$, then necessarily f must be a permutation polynomial of $\mathbb{F}_p$. By Proposition 5.3.7, the degrees $d = 2$ and $d = p-1$ are excluded, and so d satisfies $3 \leq d \leq p-2$.

Next we discuss the uniformity test for nonlinear pseudorandom numbers. If the sequence $\mathcal{S} = (x_n)_{n=0}^{\infty}$ of nonlinear pseudorandom numbers given by (5.15) is purely periodic and $\mathrm{per}(x_n) = \mathrm{per}(z_n) = p$, then the first p terms $x_0, x_1, \ldots, x_{p-1}$ of $\mathcal{S}$ run exactly through the rational numbers $0, 1/p, \ldots, (p-1)/p$ in some order. Hence in this case we get the simple formula $D_p(\mathcal{S}) = D_p^*(\mathcal{S}) = 1/p$ for the discrepancy and the star discrepancy. But for parts of the period and also for the case where

$\operatorname{per}(x_n) = \operatorname{per}(z_n) < p$, results on the (star) discrepancy of the sequence $\mathcal{S}$ are much harder to obtain. In principle, one tries to apply the same method as in Subsection 5.2.1, that is, to establish bounds on exponential sums as in Lemma 5.2.6, but now more powerful tools have to be utilized. In particular, we need the following celebrated and deep result due to André Weil (1906–1998), one of the leading mathematicians of the 20th century (see [102, Section 5.4] and [148, Section 4.4] for two different proofs of this result). Rumor has it that he found this bound while he was detained by the German occupation forces in France during World War II.

Proposition 5.3.8 (Weil Bound) *If p is a prime number and $f \in \mathbb{F}_p[x]$ is a polynomial with $\deg(f) \geq 1$, then*

$$\Big|\sum_{c\in\mathbb{F}_p} \chi_p(f(c))\Big| \leq (\deg(f)-1)p^{1/2}.$$

There is also an analog of Proposition 5.3.8 for arbitrary finite fields (see the two references above), but we do not require this general Weil bound. The following bound on exponential sums and the subsequent discrepancy bound were derived from the Weil bound by Niederreiter and Shparlinski [141], and a slight improvement was given later in [145].

Lemma 5.3.9 *Let $p \geq 3$ be a prime number and let $f \in \mathbb{F}_p[x]$ with $2 \leq d := \deg(f) < p$. Let $(z_n)_{n=0}^{\infty}$ be a purely periodic sequence of elements of $\mathbb{F}_p$ generated by the recursion (5.14). Then for every $h \in \mathbb{F}_p^*$ and every integer N with $1 \leq N \leq \operatorname{per}(z_n)$, the bound*

$$\Big|\sum_{n=0}^{N-1} \chi_p(hz_n)\Big| \leq C_1(\log d)^{1/2}N^{1/2}p^{1/2}(\log p)^{-1/2}$$

holds with an absolute constant $C_1 > 0$.

Proof. We use a method called "shift and average". We fix $h \in \mathbb{F}_p^*$ and put $\chi(c) = \chi_p(hc)$ for $c \in \mathbb{F}_p$. For every integer $m \geq 0$,

$$\sum_{n=0}^{N-1} \chi(z_n) = \sum_{n=0}^{N-1} \chi(z_{n+m}) + \theta_m$$

with $|\theta_m| \leq 2m$ since the two sums differ in at most $2m$ terms of absolute value 1. Now we choose an integer $M \geq 1$ and we sum over $m = 0, 1, \ldots, M-1$ to obtain

$$M\Big|\sum_{n=0}^{N-1} \chi(z_n)\Big| \leq W + \Big|\sum_{m=0}^{M-1} \theta_m\Big| < W + M^2 \tag{5.17}$$

with

$$W = \Big|\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\chi(z_{n+m})\Big| \le \sum_{n=0}^{N-1}\Big|\sum_{m=0}^{M-1}\chi(z_{n+m})\Big|.$$

By the Cauchy-Schwarz inequality we get

$$W^2 \le N\sum_{n=0}^{N-1}\Big|\sum_{m=0}^{M-1}\chi(z_{n+m})\Big|^2.$$

Next we introduce the polynomials $f_0, f_1, \ldots$ in $\mathbb{F}_p[x]$ by $f_0(x) = x$ and $f_m(x) = f(f_{m-1}(x))$ for $m \ge 1$. Then $z_{n+m} = f_m(z_n)$ for all $n \ge 0$ and $m \ge 0$, and so we can write

$$W^2 \le N\sum_{n=0}^{N-1}\Big|\sum_{m=0}^{M-1}\chi(f_m(z_n))\Big|^2.$$

Since the sequence $(z_n)_{n=0}^{\infty}$ is purely periodic and $N \le \mathrm{per}(z_n)$, the elements $z_0, z_1, \ldots, z_{N-1}$ of $\mathbb{F}_p$ are distinct, and therefore

$$W^2 \le N\sum_{c\in\mathbb{F}_p}\Big|\sum_{m=0}^{M-1}\chi(f_m(c))\Big|^2.$$

By expanding the square of the absolute value via $|u|^2 = u\overline{u}$ for all $u \in \mathbb{C}$, we obtain

$$W^2 \le N\sum_{c\in\mathbb{F}_p}\sum_{m,r=0}^{M-1}\chi(f_m(c) - f_r(c)) \le N\sum_{m,r=0}^{M-1}\Big|\sum_{c\in\mathbb{F}_p}\chi((f_m - f_r)(c))\Big|.$$

For the ordered pairs (m, r) with $m = r$, the inner sum has the value p and there are M such ordered pairs. For the ordered pairs (m, r) with $m \ne r$ (there are less than M^2 of these ordered pairs), we apply Proposition 5.3.8 and we note that $2 \le \deg(f_m - f_r) \le d^{M-1}$ since $\deg(f) = d \ge 2$. Therefore

$$W^2 < MNp + d^{M-1}M^2Np^{1/2}.$$

Now we choose

$$M = \left\lceil \frac{2\log p}{5\log d}\right\rceil.$$

Since $M - 1 < (2\log p)/(5\log d)$, we deduce that

$$\begin{aligned} W^2 &< \Big(\frac{2\log p}{5\log d} + 1\Big)Np + \Big(\frac{2\log p}{5\log d} + 1\Big)^2 Np^{9/10} \\ &< \frac{7\log p}{5\log d}Np + \frac{49(\log p)^2}{25(\log d)^2}Np^{9/10}. \end{aligned}$$

Now $\log p < p^{1/10}$ for sufficiently large p, and so there exists an absolute constant $C_2 > 0$ such that

$$W^2 \le C_2Np(\log p)/(\log d).$$

Together with the consequence

$$\Big|\sum_{n=0}^{N-1} \chi(z_n)\Big| < WM^{-1} + M$$

of (5.17), this leads to the final bound. □

Theorem 5.3.10 *Let $p \geq 3$ be a prime number and let $f \in \mathbb{F}_p[x]$ with $2 \leq d := \deg(f) < p$. Let $\mathcal{S} = (x_n)_{n=0}^{\infty}$ be a purely periodic sequence of nonlinear pseudorandom numbers obtained by* (5.14) *and* (5.15)*. Then for $1 \leq N \leq \mathrm{per}(x_n)$, the discrepancy bound*

$$D_N(\mathcal{S}) \leq C(\log d)^{1/2} N^{-1/2} p^{1/2} (\log p)^{-1/2} \log\log p$$

holds with an absolute constant $C > 0$.

Proof. By the Erdős-Turán inequality (see Theorem 4.1.13),

$$D_N(\mathcal{S}) \leq \frac{6}{H+1} + \frac{4}{\pi N} \sum_{h=1}^{H} \frac{1}{h} \Big| \sum_{n=0}^{N-1} \chi_p(hz_n) \Big|$$

for every integer $H \geq 1$. If also $H \leq p-1$, then we can apply Lemma 5.3.9 to obtain

$$\begin{aligned} D_N(\mathcal{S}) &\leq \frac{6}{H+1} + C_3 (\log d)^{1/2} N^{-1/2} p^{1/2} (\log p)^{-1/2} \sum_{h=1}^{H} \frac{1}{h} \\ &\leq \frac{6}{H+1} + C_3 (\log d)^{1/2} N^{-1/2} p^{1/2} (\log p)^{-1/2} (1 + \log H) \end{aligned}$$

with an absolute constant $C_3 > 0$. Now we choose

$$H = \left\lceil N^{1/2} p^{-1/2} (\log p)^{1/2} \right\rceil.$$

Then $1 \leq H \leq p-1$ and the desired bound on $D_N(\mathcal{S})$ follows immediately. □

Remark 5.3.11 You may wonder why we used the Erdős-Turán inequality in the proof of Theorem 5.3.10 and not Proposition 4.3.1 as in some other proofs of discrepancy bounds (see for instance the proof of Theorem 5.2.5). Actually, if we apply Proposition 4.3.1 with $s = 1$ and $M = p$, then we obtain

$$D_N(\mathcal{S}) \leq \frac{1}{p} + \frac{1}{N} \sum_{h \in C^*(p)} \frac{1}{r(h,p)} \Big| \sum_{n=0}^{N-1} \chi_p(hz_n) \Big|.$$

Together with Lemma 5.3.9 and (5.8), this yields

$$D_N(\mathcal{S}) \le \frac{1}{p} + C_1(\log d)^{1/2}N^{-1/2}p^{1/2}(\log p)^{-1/2}\Big(\log p + \frac{1}{3}\Big).$$

Since $N \le \mathrm{per}(z_n) \le p$, this discrepancy bound is at least of the order of magnitude $(\log p)^{1/2}$. But then this discrepancy bound is useless since $D_N(\mathcal{S}) \le 1$ is always a trivial discrepancy bound. Therefore we need the more powerful Erdős-Turán inequality in the proof of Theorem 5.3.10 in order to arrive at a nontrivial discrepancy bound. If d is small and N is of the order of magnitude p, then we have a scenario in which the bound on $D_N(\mathcal{S})$ in Theorem 5.3.10 is nontrivial.

Now we revert to the situation where $\mathrm{per}(z_n)$ attains the maximum value for fixed p, namely $\mathrm{per}(z_n) = p$. Then the map $n \in \mathbb{F}_p \mapsto z_n \in \mathbb{F}_p$ is well defined and, by Proposition 5.3.2, it can be represented by a uniquely determined polynomial $g \in \mathbb{F}_p[x]$ with $\deg(g) < p$. In other words, we can write

$$z_n = g(n) \in \mathbb{F}_p \qquad \text{for } n = 0, 1, \dots, \tag{5.18}$$

where n is also viewed as an element of $\mathbb{F}_p$. We can say a bit more since g, considered as a self-map of $\mathbb{F}_p$, must be injective if $\mathrm{per}(z_n) = p$, and so g has to be a permutation polynomial of $\mathbb{F}_p$.

This leads to the idea of the *explicit nonlinear* (*congruential*) *method* for pseudorandom number generation. Let p be a large prime number and choose a permutation polynomial g of $\mathbb{F}_p$ with $3 \le \deg(g) \le p-2$. We generate the sequence $(z_n)_{n=0}^{\infty}$ of elements of $\mathbb{F}_p$ by (5.18) and we note that $\mathrm{per}(z_n) = p$. Then we identify $\mathbb{F}_p$ with Z_p and we derive *explicit nonlinear* (*congruential*) *pseudorandom numbers* $x_0, x_1, \dots$ by (5.15).

If $\mathcal{S} = (x_n)_{n=0}^{\infty}$ is a sequence of explicit nonlinear pseudorandom numbers, then it follows from $\mathrm{per}(x_n) = \mathrm{per}(z_n) = p$ that the first p terms $x_0, x_1, \dots, x_{p-1}$ of $\mathcal{S}$ run exactly through the rational numbers $0, 1/p, \dots, (p-1)/p$ in some order. Therefore as in an earlier case we get $D_p(\mathcal{S}) = D_p^*(\mathcal{S}) = 1/p$ for the discrepancy and the star discrepancy. For parts of the period, we apply a method that is similar to that for linear congruential pseudorandom numbers (see Lemma 5.2.6 and Theorem 5.2.7).

Lemma 5.3.12 *If p is a prime number and g is a polynomial over $\mathbb{F}_p$ with $\deg(g) \ge 2$, then*

$$\Big|\sum_{n=0}^{N-1} \chi_p(g(n))\Big| < (\deg(g)-1)p^{1/2}\Big(\log p + \frac{4}{3}\Big)$$

for all integers N with $1 \le N < p$. If g is a permutation polynomial of $\mathbb{F}_p$ with $\deg(g) \ge 2$, then this bound can be slightly improved to

$$\Big|\sum_{n=0}^{N-1} \chi_p(g(n))\Big| < (\deg(g)-1)p^{1/2}\Big(\log p + \frac{1}{3}\Big) \quad \textit{for } 1 \le N < p.$$

Proof. We can assume that $p \geq 3$. We start from an obvious analog of the identity (5.9), namely

$$\sum_{n=0}^{N-1} \chi_p(g(n)) = \sum_{n=0}^{p-1} \chi_p(g(n)) \sum_{r=0}^{N-1} \frac{1}{p} \sum_{c=0}^{p-1} \chi_p(c(n-r)).$$

We rewrite this identity in the form

$$\sum_{n=0}^{N-1} \chi_p(g(n)) = \frac{1}{p} \sum_{c=0}^{p-1} \Big(\sum_{r=0}^{N-1} \chi_p(-cr) \Big) \Big(\sum_{n=0}^{p-1} \chi_p(g(n) + cn) \Big).$$

By taking absolute values, we get

$$\Big| \sum_{n=0}^{N-1} \chi_p(g(n)) \Big| \leq \frac{1}{p} \sum_{c=0}^{p-1} \Big| \sum_{r=0}^{N-1} \chi_p(cr) \Big| \Big| \sum_{n=0}^{p-1} \chi_p(g(n) + cn) \Big|.$$

For all $c \in \mathbb{F}_p$, the polynomial $f(x) = g(x) + cx \in \mathbb{F}_p[x]$ satisfies $\deg(f) = \deg(g) \geq 2$, and so we can apply Proposition 5.3.8 to the last exponential sum to obtain

$$\Big| \sum_{n=0}^{N-1} \chi_p(g(n)) \Big| \leq (\deg(g) - 1) p^{-1/2} \sum_{c=0}^{p-1} \Big| \sum_{r=0}^{N-1} \chi_p(cr) \Big|.$$

Finally, by proceeding as in the proof of Lemma 5.2.6, we get

$$\sum_{c=0}^{p-1} \Big| \sum_{r=0}^{N-1} \chi_p(cr) \Big| = N + \sum_{c=1}^{p-1} \Big| \sum_{r=0}^{N-1} \chi_p(cr) \Big| < p \Big(\log p + \frac{4}{3} \Big),$$

and so we arrive at the first bound in the lemma.

In order to obtain the second bound in the lemma, we simply note that for $c = 0$ the identity

$$\sum_{n=0}^{p-1} \chi_p(g(n) + cn) = \sum_{n=0}^{p-1} \chi_p(g(n)) = \sum_{n=0}^{p-1} \chi_p(n) = 0$$

holds whenever g is a permutation polynomial of $\mathbb{F}_p$. □

Theorem 5.3.13 *Let $p \geq 5$ be a prime number and let $g \in \mathbb{F}_p[x]$ be a permutation polynomial of $\mathbb{F}_p$ with $3 \leq \deg(g) \leq p - 2$. Let $\mathcal{S} = (x_n)_{n=0}^{\infty}$ be the sequence of explicit nonlinear pseudorandom numbers obtained by (5.18) and (5.15). Then the discrepancy bound*

$$D_N(\mathcal{S}) < (\deg(g) - 1) N^{-1} p^{1/2} \Big(\log p + \frac{1}{3} \Big)^2 + \frac{1}{p}$$

is valid for all integers N with $1 \leq N < p$.

Proof. From Proposition 4.3.1 with $s = 1$ and $M = p$ we get

$$D_N(\mathcal{S}) \le \frac{1}{p} + \frac{1}{N} \sum_{h \in C^*(p)} \frac{1}{r(h,p)} \Big| \sum_{n=0}^{N-1} \chi_p(hg(n)) \Big|.$$

Now we apply Lemma 5.3.12 to obtain

$$D_N(\mathcal{S}) < \frac{1}{p} + (\deg(g) - 1) N^{-1} p^{1/2} \Big(\log p + \frac{1}{3} \Big) \sum_{h \in C^*(p)} \frac{1}{r(h,p)}.$$

We conclude the proof by invoking the inequality (5.8). □

We deduce from Theorem 5.3.13 that if the degree of the polynomial g is small compared to p and if N is of the order of magnitude p, then the upper bound on $D_N(\mathcal{S})$, and therefore on $D_N^*(\mathcal{S})$, is of the order of magnitude $N^{-1/2}(\log N)^2$. This is in reasonably good accordance with the law of the iterated logarithm for the star discrepancy (see Subsection 5.1.2). Results analogous to Theorem 5.3.13 can be established also for the serial test for explicit nonlinear pseudorandom numbers (see [132] and Exercises 5.13 and 5.14).

5.3.2 Inversive methods

The discrepancy bound for nonlinear pseudorandom numbers shown in Theorem 5.3.10 is nontrivial, but nevertheless very weak. The poor quality of this result stems from the nature of the recursion (5.14) and from the fact that if we iterate a polynomial f over $\mathbb{F}_p$ with $\deg(f) \ge 2$, then the degrees of the iterates grow exponentially. We can remedy this situation if we can find interesting functions on $\mathbb{F}_p$ that do not exhibit this phenomenon of the explosion of degrees under iteration. Such a family of functions is given by linear fractional transformations on $\mathbb{F}_p$ which are rational functions of the form $r(x) = (a_1 + b_1 x)/(a_2 + b_2 x)$ with $a_1, a_2, b_1, b_2 \in \mathbb{F}_p$ and $a_1 b_2 - a_2 b_1 \ne 0$. An easy computation shows that the composition of two linear fractional transformations on $\mathbb{F}_p$ is again a linear fractional transformation on $\mathbb{F}_p$. We are thus led to consider the recursion (5.14) with the polynomial $f(x)$ replaced by the rational function $r(x)$. In order to avoid a linear recursion, we assume that $b_2 \ne 0$. Then with a linear substitution, we can simplify the form of $r(x)$ to $r(x) = (a + bx)/x = ax^{-1} + b$ with $a \ne 0$.

Now we come to the formal definition of this method which, since it utilizes multiplicative inverses in $\mathbb{F}_p$, is called the *inversive* (*congruential*) *method*. We stay away from trivial cases by taking a prime number $p \ge 5$, but in practice p will of course be a large prime number such as $p = 2^{31} - 1$. We choose $a, b \in \mathbb{F}_p$ with $a \ne 0$ and consider the recursion $z_{n+1} = r(z_n)$ for $n = 0, 1, \ldots$ with $r(x) = ax^{-1} + b$. There is a slight technical problem here since $r(0)$ is not defined. But we may introduce a sort of pseudo-inverse of $0 \in \mathbb{F}_p$ by brute force: it is reasonable to define the pseudo-inverse of 0 to be 0 since this is the only element of $\mathbb{F}_p$ that does not show up as a multiplicative inverse of an element of $\mathbb{F}_p^*$. Concretely, for every $c \in \mathbb{F}_p$ we introduce the notation

$$\overline{c} = \begin{cases} c^{-1} \in \mathbb{F}_p & \text{if } c \neq 0, \\ 0 \in \mathbb{F}_p & \text{if } c = 0. \end{cases}$$

Finally, we now generate a sequence $(z_n)_{n=0}^{\infty}$ of elements of $\mathbb{F}_p$ by choosing an initial value $z_0 \in \mathbb{F}_p$ and proceeding by the recursion

$$z_{n+1} = a\overline{z_n} + b \qquad \text{for } n = 0, 1, \ldots. \tag{5.19}$$

Then we identify $\mathbb{F}_p$ with Z_p and we obtain *inversive (congruential) pseudorandom numbers* by the normalization

$$x_n = \frac{z_n}{p} \in [0, 1) \qquad \text{for } n = 0, 1, \ldots. \tag{5.20}$$

These pseudorandom numbers were proposed by Eichenauer and Lehn [43], even before the general nonlinear method was defined. Since z_n is uniquely determined by z_{n+1} in (5.19) in view of $a \neq 0$, the sequence $(z_n)_{n=0}^{\infty}$ is purely periodic, and so is the sequence $(x_n)_{n=0}^{\infty}$. As in the general nonlinear method, it is evident that $\mathrm{per}(x_n) = \mathrm{per}(z_n) \leq p$.

Remark 5.3.14 If you prefer, you can write $\overline{c} = c^{p-2}$ for all $c \in \mathbb{F}_p$. This is trivial for $c = 0$, whereas for $c \neq 0$ we observe that Proposition 1.4.13 yields $1 = c^{p-1} = c(c^{p-2})$, and so $\overline{c} = c^{-1} = c^{p-2}$. However, in the proofs it will be more useful to think of $\overline{c}$ as being basically the multiplicative inverse of c.

There are choices of the parameters a and b in (5.19) that yield small period lengths. A really bad choice is $b = 0$, for then it is readily seen that $z_{n+2} = z_n$ for $n = 0, 1, \ldots$, and so $\mathrm{per}(x_n) = \mathrm{per}(z_n) \leq 2$. On the other hand, we can always select a and b in such a way that we get the theoretically largest possible least period length $\mathrm{per}(x_n) = \mathrm{per}(z_n) = p$ for fixed p, as we shall see below. The property $\mathrm{per}(z_n) = p$ is connected with the roots α and β of the polynomial $f(x) = x^2 - bx - a \in \mathbb{F}_p[x]$. Since $\deg(f) = 2$, the roots α and β lie in the extension field $\mathbb{F}_{p^2}$ of $\mathbb{F}_p$. Furthermore, $x^2 - bx - a = (x - \alpha)(x - \beta)$ implies that $\alpha\beta = -a \neq 0$, and so $\alpha\beta^{-1} \in \mathbb{F}_{p^2}^*$.

Lemma 5.3.15 *Let $\alpha, \beta \in \mathbb{F}_{p^2}$ be the roots of $f(x) = x^2 - bx - a \in \mathbb{F}_p[x]$ and assume that $\alpha \neq \beta$. Let k be the order of $\alpha\beta^{-1}$ in the multiplicative group $\mathbb{F}_{p^2}^*$. Then the sequence $(z_n)_{n=0}^{\infty}$ generated by (5.19) with the initial value $z_0 = b$ satisfies $\mathrm{per}(z_n) = k - 1$.*

Proof. By the definition of k we can say that $\alpha^n \neq \beta^n$ for $1 \leq n \leq k-1$ and $\alpha^k = \beta^k$. We claim that

$$z_n = \frac{\alpha^{n+2} - \beta^{n+2}}{\alpha^{n+1} - \beta^{n+1}} \qquad \text{for } n = 0, 1, \ldots, k-2. \tag{5.21}$$

For $n = 0$ we get $(\alpha^2 - \beta^2)/(\alpha - \beta) = \alpha + \beta = b = z_0$, and so (5.21) holds. Suppose that (5.21) is shown for some n with $0 \leq n \leq k-3$. Then $z_n \neq 0$ and the recursion (5.19) yields

$$\begin{aligned} z_{n+1} &= a z_n^{-1} + b = a\,\frac{\alpha^{n+1} - \beta^{n+1}}{\alpha^{n+2} - \beta^{n+2}} + b \\ &= \frac{-\alpha\beta(\alpha^{n+1} - \beta^{n+1}) + (\alpha+\beta)(\alpha^{n+2} - \beta^{n+2})}{\alpha^{n+2} - \beta^{n+2}} = \frac{\alpha^{n+3} - \beta^{n+3}}{\alpha^{n+2} - \beta^{n+2}}. \end{aligned}$$

Hence the proof of (5.21) by induction is complete. Now (5.21) implies that $z_n \neq 0$ for $0 \leq n \leq k-3$ and $z_{k-2} = 0$. Then as a consequence of (5.19) we get $z_n \neq b$ for $1 \leq n \leq k-2$ and $z_{k-1} = b$. Therefore $\mathrm{per}(z_n) = k-1$ since the sequence $(z_n)_{n=0}^{\infty}$ is purely periodic. □

Lemma 5.3.15 is the crucial step in the proof of the following theorem that provides an attractive criterion for the property $\mathrm{per}(x_n) = \mathrm{per}(z_n) = p$.

Theorem 5.3.16 *Let $p \geq 5$ be a prime number and let $a, b \in \mathbb{F}_p$ with $a \neq 0$. Let $\alpha, \beta \in \mathbb{F}_{p^2}$ be the roots of $f(x) = x^2 - bx - a \in \mathbb{F}_p[x]$. Then a sequence $(z_n)_{n=0}^{\infty}$ generated by (5.19) satisfies $\mathrm{per}(z_n) = p$ if and only if the order of $\alpha\beta^{-1}$ in the multiplicative group $\mathbb{F}_{p^2}^*$ is equal to $p+1$.*

Proof. Let us first analyze the degenerate case where $\alpha = \beta$. Then $a = -\alpha^2$ and $b = 2\alpha$. If we had $\mathrm{per}(z_n) = p$, then $z_m = \alpha$ for some $m \geq 0$. But then (5.19) yields $z_{m+1} = a z_m^{-1} + b = -\alpha^2\alpha^{-1} + 2\alpha = \alpha = z_m$, a contradiction. Hence in this case we have $\mathrm{per}(z_n) < p$ and also the order of $\alpha\beta^{-1} = 1$ in $\mathbb{F}_{p^2}^*$ is 1 and not $p+1$.

Thus, it remains to treat the case where $\alpha \neq \beta$. If $\mathrm{per}(z_n) = p$, then $z_h = b$ for some $h \geq 0$. Now we consider the shifted sequence $(y_n)_{n=0}^{\infty}$ defined by $y_n = z_{n+h}$ for all $n \geq 0$. Then $(y_n)_{n=0}^{\infty}$ satisfies the recursion (5.19) as well as $y_0 = b$ and $\mathrm{per}(y_n) = p$. Hence it follows from Lemma 5.3.15 that the order k of $\alpha\beta^{-1}$ in $\mathbb{F}_{p^2}^*$ satisfies $k - 1 = \mathrm{per}(y_n) = p$, and so $k = p+1$. Conversely, suppose that the order of $\alpha\beta^{-1}$ in $\mathbb{F}_{p^2}^*$ is $p+1$. Let $(w_n)_{n=0}^{\infty}$ be the sequence generated by (5.19) with the initial value $w_0 = b$. Then $\mathrm{per}(w_n) = p$ by Lemma 5.3.15. Therefore $w_j = z_0$ for some $j \geq 0$, hence the sequence $(z_n)_{n=0}^{\infty}$ is a shifted version of $(w_n)_{n=0}^{\infty}$, and so also $\mathrm{per}(z_n) = p$. □

Example 5.3.17 Let $f(x) = x^2 - bx - a \in \mathbb{F}_p[x]$ be a primitive quadratic polynomial over $\mathbb{F}_p$. Then the roots of f are $\beta \in \mathbb{F}_{p^2}^*$ and $\alpha = \beta^p$ (see Proposition 1.4.47) and β is a primitive element of $\mathbb{F}_{p^2}$. Now $\alpha\beta^{-1} = \beta^{p-1}$ has order $p+1$ in the multiplicative group $\mathbb{F}_{p^2}^*$ since β has order $p^2 - 1$ in $\mathbb{F}_{p^2}^*$. Thus, if $a, b \in \mathbb{F}_p$ are chosen in such a way that $x^2 - bx - a$ is a primitive quadratic polynomial over $\mathbb{F}_p$, then Theorem 5.3.16 shows that every sequence $(z_n)_{n=0}^{\infty}$ generated by (5.19) with these values of a and b satisfies $\operatorname{per}(z_n) = p$. Note that a primitive quadratic polynomial over $\mathbb{F}_p$ exists for every prime number p (see Proposition 1.4.43).

As we found out in Subsection 5.2.2, the points of the point set $\mathcal{P}^*(\mathbf{g}(a), p)$ obtained from linear congruential pseudorandom numbers “fall mainly in the planes”, and this can have devastating effects on Monte Carlo computations with these points (see Examples 5.2.12 and 5.3.1). It is a splendid feature of inversive pseudorandom numbers that the corresponding points derived from them exhibit the contrary behavior and “avoid the planes”, in the fitting words of the paper [44]. The setting for this result is the vector space $\mathbb{F}_p^s$ of dimension $s \geq 2$. Just as in the Euclidean space $\mathbb{R}^s$, we can talk about hyperplanes in $\mathbb{F}_p^s$; namely, a hyperplane in $\mathbb{F}_p^s$ is a set of the form $H = \{\mathbf{v} \in \mathbb{F}_p^s : \mathbf{h} \cdot \mathbf{v} = c\}$ with a fixed nonzero vector $\mathbf{h} \in \mathbb{F}_p^s$ and a fixed $c \in \mathbb{F}_p$.

Theorem 5.3.18 *Let $p \geq 5$ be a prime number and let $s \geq 2$ be an integer. Let $(z_n)_{n=0}^{\infty}$ be a sequence generated by* (5.19) *with* $\operatorname{per}(z_n) = p$. *Then every hyperplane in $\mathbb{F}_p^s$ contains at most s of the points*

$$\mathbf{z}_n = (z_n, z_{n+1}, \ldots, z_{n+s-1}) \in \mathbb{F}_p^s$$

with $n = 0, 1, \ldots, p-1$ and $z_n \cdots z_{n+s-2} \neq 0$.

Proof. Let $(y_n)_{n=0}^{\infty}$ be the sequence generated by (5.19) with the initial value $y_0 = 0$. Then $(y_n)_{n=0}^{\infty}$ is a shifted version of $(z_n)_{n=0}^{\infty}$, and so $\operatorname{per}(y_n) = p$. We put $d_j = -a\overline{y_j} \in \mathbb{F}_p$ for $j \geq 0$. It follows from $\{y_0, y_1, \ldots, y_{p-1}\} = \mathbb{F}_p$ that $d_0, d_1, \ldots, d_{p-1}$ are distinct. Define $\psi(n) = a\overline{n} + b \in \mathbb{F}_p$ for $n \in \mathbb{F}_p$ and let ψ^j be the jth iterate of the map ψ, with ψ^0 being the identity map on $\mathbb{F}_p$. By a straightforward induction on j it is proved that

$$\psi^j(n) = y_j \frac{n - d_j}{n - d_{j-1}} \qquad \text{for } 1 \leq j \leq p-1 \tag{5.22}$$

and whenever $n \neq d_i$ for $0 \leq i \leq j-1$. Since the theorem is trivial for $s \geq p$, we can assume that $s < p$. From $\operatorname{per}(z_n) = p$ we infer that

$$\{\mathbf{z}_n : 0 \leq n \leq p-1\} = \{\big(\psi^0(n), \psi^1(n), \ldots, \psi^{s-1}(n)\big) : 0 \leq n \leq p-1\}. \tag{5.23}$$

It follows from (5.22) that the condition $z_n \cdots z_{n+s-2} \neq 0$ amounts to the condition $n \neq d_i$ for $0 \leq i \leq s-2$ in the second set in (5.23).

Now let a hyperplane H in $\mathbb{F}_p^s$ be given. Concretely, let $H = \{\mathbf{v} \in \mathbb{F}_p^s : \mathbf{h} \cdot \mathbf{v} = c\}$ with a fixed nonzero $\mathbf{h} = (h_1, \ldots, h_s) \in \mathbb{F}_p^s$ and a fixed $c \in \mathbb{F}_p$. Then (5.22) shows that if $n \neq d_i$ for $0 \leq i \leq s-2$, then $(\psi^0(n), \psi^1(n), \ldots, \psi^{s-1}(n)) \in H$ if and only if

$$h_1 n + \sum_{j=2}^{s} h_j y_{j-1} \frac{n - d_{j-1}}{n - d_{j-2}} = c.$$

Clearing denominators, we see that this is equivalent to $g(n) = 0$, where the polynomial $g \in \mathbb{F}_p[x]$ is given by

$$g(x) = (h_1 x - c) \prod_{j=2}^{s} (x - d_{j-2}) + \sum_{j=2}^{s} h_j y_{j-1} (x - d_{j-1}) \prod_{\substack{i=2 \\ i \neq j}}^{s} (x - d_{i-2}).$$

Since $\deg(g) \leq s$, the final result follows from Theorem 1.4.27 if we can verify that g is not the zero polynomial. If g were the zero polynomial, then by considering the coefficient of x^s we would get $h_1 = 0$. Furthermore, for $2 \leq k \leq s$ we would obtain

$$0 = g(d_{k-2}) = h_k y_{k-1} (d_{k-2} - d_{k-1}) \prod_{\substack{i=2 \\ i \neq k}}^{s} (d_{k-2} - d_{i-2}).$$

We know that all factors on the right-hand side except possibly h_k are nonzero, and so we would get $h_k = 0$. This is a contradiction to $(h_1, \ldots, h_s) \neq \mathbf{0}$. □

Remark 5.3.19 As in the Euclidean space $\mathbb{R}^s$, for any s given points in $\mathbb{F}_p^s$ there is a hyperplane passing through these points. Therefore Theorem 5.3.18 is optimal, in the sense that for sufficiently large p there do exist hyperplanes in $\mathbb{F}_p^s$ that contain exactly s of the points $\mathbf{z}_n$ considered in this theorem. Since $\{z_0, z_1, \ldots, z_{p-1}\} = \mathbb{F}_p$ and $\mathrm{per}(z_n) = p$, the condition $z_n \cdots z_{n+s-2} \neq 0$ eliminates exactly $s-1$ of the points $\mathbf{z}_n$ in the range $0 \leq n \leq p-1$.

Remark 5.3.20 Let us check what happens in Theorem 5.3.18 if we replace the sequence $(z_n)_{n=0}^{\infty}$ there by a sequence obtained from the linear congruential method with the prime modulus p. Then (5.4) shows that $z_n = a^n z_0$ in $\mathbb{F}_p$ for $n = 0, 1, \ldots$, where $a \in \mathbb{F}_p^*$ and $z_0 \in \mathbb{F}_p^*$. Consequently, for every $s \geq 2$ we get

$$\mathbf{z}_n = (z_n, z_{n+1}, \ldots, z_{n+s-1}) = \big(a^n z_0, a^{n+1} z_0, \ldots, a^{n+s-1} z_0\big) \in \mathbb{F}_p^s$$

for $n = 0, 1, \ldots$. It follows that *all* points $\mathbf{z}_n$ lie in the hyperplane $H = \{\mathbf{v} \in \mathbb{F}_p^s : \mathbf{h} \cdot \mathbf{v} = 0\}$ in $\mathbb{F}_p^s$ with $\mathbf{h} = (a, -1, 0, \ldots, 0) \in \mathbb{F}_p^s$. The lesson is that in terms of structural properties, inversive pseudorandom numbers are vastly superior to linear congruential pseudorandom numbers.

Now we turn to the uniformity test for inversive pseudorandom numbers. Let $\mathcal{S} = (x_n)_{n=0}^{\infty}$ be a sequence of inversive pseudorandom numbers given by (5.19) and (5.20), and we again focus on the case where $\mathrm{per}(x_n) = \mathrm{per}(z_n) = p$. Then as for the general nonlinear method in Subsection 5.3.1, we get $D_p(\mathcal{S}) = D_p^*(\mathcal{S}) = 1/p$ for the discrepancy and the star discrepancy. For parts of the period, we could try to utilize the discrepancy bound for the general nonlinear method in Theorem 5.3.10. Note that the inversive method is the nonlinear method with the special polynomial $f(x) = ax^{p-2} + b \in \mathbb{F}_p[x]$ (see Remark 5.3.14). But then $d = p - 2$ in the notation of Theorem 5.3.10, and so $\log d$ has the order of magnitude $\log p$. It follows that the discrepancy bound in Theorem 5.3.10 is at least of the order of magnitude $\log\log p$, and so it is useless.

In fact, it was an open problem for many years to prove a nontrivial bound on the discrepancy $D_N(\mathcal{S})$ for $1 \le N < p$. This was finally achieved by Niederreiter and Shparlinski in the paper [142]. As a technical tool, we need a bound for a classical family of exponential sums called Kloosterman sums. We state this bound in the following proposition and we refer to [102, Section 5.5] for a proof.

Proposition 5.3.21 *Let p be a prime number and suppose that $a_1, a_2 \in \mathbb{F}_p$ are not both 0. Then*

$$\Big| \sum_{c \in \mathbb{F}_p^*} \chi_p(a_1 c^{-1} + a_2 c) \Big| \le 2p^{1/2}.$$

Lemma 5.3.22 *Let $p \ge 5$ be a prime number, let $(z_n)_{n=0}^{\infty}$ be a sequence of elements of $\mathbb{F}_p$ generated by the recursion (5.19) with $\mathrm{per}(z_n) = p$, and let $h \in \mathbb{F}_p^*$. Then*

$$\Big| \sum_{n=0}^{N-1} \chi_p(hz_n) \Big| < 3N^{1/2}p^{1/4} + (3p/2)^{1/2} \tag{5.24}$$

for all integers N with $1 \le N < p$.

Proof. We again use the method "shift and average" that we already employed in the proof of Lemma 5.3.9. Since the right-hand side of (5.24) is greater than 5 for all $N \ge 1$ and $p \ge 5$, we can assume that $p \ge 7$. We fix $h \in \mathbb{F}_p^*$ and put $\chi(c) = \chi_p(hc)$ for all $c \in \mathbb{F}_p$. By repeating the argument in the beginning of the proof of Lemma 5.3.9, we get

$$\Big| \sum_{n=0}^{N-1} \chi(z_n) \Big| < WM^{-1} + M \tag{5.25}$$

and

$$W^2 \le N \sum_{n=0}^{N-1} \Big| \sum_{m=0}^{M-1} \chi(z_{n+m}) \Big|^2,$$

where M is an arbitrary positive integer. Let $\psi : \mathbb{F}_p \to \mathbb{F}_p$ be defined by $\psi(c) = a\bar{c} + b$ for all $c \in \mathbb{F}_p$. Then (5.19) implies that $z_{n+m} = \psi^m(z_n)$ for all $n \ge 0$ and $m \ge 0$, where ψ^m is the mth iterate of ψ (compare with the proof of Theorem 5.3.18). Hence we obtain

$$W^2 \le N \sum_{n=0}^{N-1} \Big| \sum_{m=0}^{M-1} \chi(\psi^m(z_n)) \Big|^2 .$$

Since $N < p = \mathrm{per}(z_n)$, the elements $z_0, z_1, \dots, z_{N-1}$ of $\mathbb{F}_p$ are distinct, and so

$$W^2 \le N \sum_{c \in \mathbb{F}_p} \Big| \sum_{m=0}^{M-1} \chi(\psi^m(c)) \Big|^2 .$$

By expanding the square of the absolute value, we get

$$\begin{aligned} W^2 &\le N \sum_{c \in \mathbb{F}_p} \sum_{m,r=0}^{M-1} \chi(\psi^m(c) - \chi^r(c)) \\ &\le N \sum_{m,r=0}^{M-1} \Big| \sum_{c \in \mathbb{F}_p} \chi(\psi^m(c) - \psi^r(c)) \Big| \\ &= MNp + 2N \sum_{\substack{m,r=0 \\ m>r}}^{M-1} \Big| \sum_{c \in \mathbb{F}_p} \chi(\psi^m(c) - \psi^r(c)) \Big| . \end{aligned}$$

The last exponential sum can be written in the form

$$\sum_{c \in \mathbb{F}_p} \chi(\psi^m(c) - \psi^r(c)) = \sum_{c \in \mathbb{F}_p} \chi(\psi^{m-r}(\psi^r(c)) - \psi^r(c)).$$

Since ψ^r is a permutation of $\mathbb{F}_p$, we can take $\psi^r(c)$ as a new summation variable over $\mathbb{F}_p$, and this yields

$$\sum_{c \in \mathbb{F}_p} \chi(\psi^m(c) - \psi^r(c)) = \sum_{c \in \mathbb{F}_p} \chi(\psi^{m-r}(c) - c).$$

It is therefore reasonable to combine the contributions of all ordered pairs (m, r) with fixed difference $m - r = k \ge 1$. There are $M - k$ such ordered pairs in the given range for m and r, and so we arrive at the inequality

$$W^2 \le MNp + 2N \sum_{k=1}^{M-1} (M-k) \Big| \sum_{c \in \mathbb{F}_p} \chi(\psi^k(c) - c) \Big| . \tag{5.26}$$

Now we study the last exponential sum for a fixed k with $1 \le k \le M-1$ and we assume that $M \le p$. We can then apply the formula (5.22) to $\psi^k(c)$ as long as $c \notin E_k := \{d_0, d_1, \dots, d_{k-1}\}$. Since E_k has k elements, we obtain

$$\begin{aligned} \Big| \sum_{c \in \mathbb{F}_p} \chi(\psi^k(c) - c) \Big| &\le \Big| \sum_{c \in \mathbb{F}_p \setminus E_k} \chi\Big(y_k \frac{c - d_k}{c - d_{k-1}} - c \Big) \Big| + k \\ &\le \Big| \sum_{c \in \mathbb{F}_p \setminus \{d_{k-1}\}} \chi\Big(y_k \frac{c - d_k}{c - d_{k-1}} - c \Big) \Big| + 2k - 1. \end{aligned}$$

In the last exponential sum we introduce $w = c - d_{k-1}$ as a new summation variable. This yields

$$\begin{aligned}\Big|\sum_{c\in\mathbb{F}_p\setminus\{d_{k-1}\}} \chi\Big(y_k\,\frac{c-d_k}{c-d_{k-1}} - c\Big)\Big| &= \Big|\sum_{w\in\mathbb{F}_p^*} \chi\Big(y_k\,\frac{w+d_{k-1}-d_k}{w} - w - d_{k-1}\Big)\Big| \\ &= \Big|\sum_{w\in\mathbb{F}_p^*} \chi(y_k(d_{k-1}-d_k)w^{-1} - w)\Big|.\end{aligned}$$

The last exponential sum is a Kloosterman sum, and so we can apply Proposition 5.3.21 to obtain

$$\Big|\sum_{c\in\mathbb{F}_p} \chi(\psi^k(c) - c)\Big| \le 2p^{1/2} + 2k - 1.$$

By plugging this bound into (5.26), we get

$$W^2 \le MNp + 2N\sum_{k=1}^{M-1}(M-k)(2p^{1/2}+2k-1).$$

A straightforward computation shows that

$$\sum_{k=1}^{M-1}(M-k)(2k-1) = M(M-1)\Big(\frac{M}{3}-\frac{1}{6}\Big) < \frac{1}{3}M^3.$$

Thus, we arrive at the bound

$$W^2 < MNp + 2M^2Np^{1/2} + \frac{2}{3}M^3N.$$

Now we put $M = \lfloor (3p/2)^{1/2} \rfloor$ which is a permitted value since obviously $1 \le M \le p$. With this choice for M we get $W^2 < (3+\sqrt{6})Np^{3/2}$. Recalling that $p \ge 7$, we deduce that

$$\frac{W}{M} < \frac{(3+\sqrt{6})^{1/2}N^{1/2}p^{3/4}}{(3p/2)^{1/2}-1} = \frac{(3+\sqrt{6})^{1/2}N^{1/2}p^{1/4}}{\sqrt{3/2}-p^{-1/2}} \le \frac{(3+\sqrt{6})^{1/2}}{\sqrt{3/2}-\sqrt{1/7}}N^{1/2}p^{1/4}.$$

After having some fun with computing square roots, we get $WM^{-1} < 3N^{1/2}p^{1/4}$, and in view of (5.25) we arrive at the bound in (5.24). □

Theorem 5.3.23 *Let $p \ge 5$ be a prime number and let $\mathcal{S} = (x_n)_{n=0}^{\infty}$ be a sequence of inversive pseudorandom numbers obtained by (5.19) and (5.20) with* $\mathrm{per}(x_n) = p$. *Then the discrepancy bound*

$$D_N(\mathcal{S}) < \big(3N^{-1/2}p^{1/4} + (3p/2)^{1/2}N^{-1}\big)\Big(\log p + \frac{1}{3}\Big) + \frac{1}{p}$$

holds for all integers N with $1 \le N < p$.

Proof. This bound is derived from Lemma 5.3.22 in the same way as Theorem 5.3.13 was derived from Lemma 5.3.12. □

The discrepancy bound in Theorem 5.3.23 is nontrivial as soon as N is somewhat larger than $p^{1/2}$, say at least of the order of magnitude $p^{(1/2)+\varepsilon}$ with an $\varepsilon > 0$ independent of p. If N has the order of magnitude p, then $D_N(\mathcal{S}) = O\left(p^{-1/4}\log p\right)$ with an absolute implied constant. It is an open problem whether this can be improved to $D_N(\mathcal{S}) = O\left(p^{-1/2}(\log p)^c\right)$ for some absolute constant $c \geq 0$. Results on the serial test for inversive pseudorandom numbers are summarized in the survey article of Niederreiter and Shparlinski [143].

For the general nonlinear method, we discussed an explicit counterpart to the recursive procedure for generating pseudorandom numbers (see Subsection 5.3.1). There is also an explicit version of the inversive method which was proposed by Eichenauer-Herrmann [45]. By the way, Eichenauer and Eichenauer-Herrmann is the same person, before and after marriage. Let $p \geq 5$ be a prime number. We recall the notation $\overline{c} \in \mathbb{F}_p$ for $c \in \mathbb{F}_p$ which stands for $\overline{c} = 0$ if $c = 0$ and $\overline{c} = c^{-1}$ if $c \in \mathbb{F}_p^*$. Now we choose $a, b \in \mathbb{F}_p$ with $a \neq 0$ and we generate the sequence $(z_n)_{n=0}^{\infty}$ by the explicit formula

$$z_n = \overline{an + b} \in \mathbb{F}_p \qquad \text{for } n = 0, 1, \ldots . \tag{5.27}$$

Then we identify $\mathbb{F}_p$ with Z_p and we obtain *explicit inversive* (*congruential*) *pseudorandom numbers* by putting

$$x_n = \frac{z_n}{p} \in [0, 1) \qquad \text{for } n = 0, 1, \ldots . \tag{5.28}$$

It is obvious that $\mathrm{per}(x_n) = \mathrm{per}(z_n) = p$.

The analysis of explicit inversive pseudorandom numbers is much easier than that of the inversive pseudorandom numbers generated by (5.19) and (5.20). If $\mathcal{S} = (x_n)_{n=0}^{\infty}$ is a sequence of explicit inversive pseudorandom numbers, then $D_p(\mathcal{S}) = D_p^*(\mathcal{S}) = 1/p$ as in earlier cases. For parts of the period, we proceed by a method that we have already utilized before.

Lemma 5.3.24 *Let $p \geq 5$ be a prime number, let $(z_n)_{n=0}^{\infty}$ be a sequence of elements of $\mathbb{F}_p$ defined by* (5.27) *with $a, b \in \mathbb{F}_p$ and $a \neq 0$, and let $h \in \mathbb{F}_p^*$. Then*

$$\Big|\sum_{n=0}^{N-1} \chi_p(hz_n)\Big| < \left(2p^{1/2} + 1\right)\left(\log p + \frac{1}{3}\right) \qquad \text{for } 1 \leq N < p.$$

Proof. As in the proof of Lemma 5.3.12 we can write

$$\begin{aligned}\sum_{n=0}^{N-1} \chi_p(hz_n) &= \sum_{n=0}^{p-1} \chi_p(hz_n) \sum_{r=0}^{N-1} \frac{1}{p} \sum_{d=0}^{p-1} \chi_p(d(n-r)) \\ &= \frac{1}{p}\sum_{d=0}^{p-1}\Big(\sum_{r=0}^{N-1}\chi_p(-dr)\Big)\Big(\sum_{n=0}^{p-1}\chi_p(hz_n + dn)\Big).\end{aligned}$$

Therefore

$$\Big|\sum_{n=0}^{p-1} \chi_p(hz_n)\Big| \leq \frac{1}{p}\sum_{d=0}^{p-1}\Big|\sum_{r=0}^{N-1}\chi_p(dr)\Big|\Big|\sum_{n\in\mathbb{F}_p}\chi_p(h\,\overline{an+b}+dn)\Big|.$$

We can take the outer sum from $d = 1$ to $d = p-1$ since the contribution for $d = 0$ is equal to 0. With the substitution $c = an+b \in \mathbb{F}_p$ in the last exponential sum, we obtain

$$\begin{aligned}\Big|\sum_{n\in\mathbb{F}_p}\chi_p(h\,\overline{an+b}+dn)\Big| &= \Big|\sum_{c\in\mathbb{F}_p}\chi_p(h\overline{c}+da^{-1}(c-b))\Big| \\ &= \Big|\sum_{c\in\mathbb{F}_p}\chi_p(h\overline{c}+da^{-1}c)\Big|.\end{aligned}$$

Now an application of Proposition 5.3.21 yields

$$\Big|\sum_{c\in\mathbb{F}_p}\chi_p(h\overline{c}+da^{-1}c)\Big| \leq 1+\Big|\sum_{c\in\mathbb{F}_p^*}\chi_p(hc^{-1}+da^{-1}c)\Big| \leq 1+2p^{1/2}.$$

Therefore

$$\Big|\sum_{n=0}^{N-1}\chi_p(hz_n)\Big| \leq \frac{2p^{1/2}+1}{p}\sum_{d=1}^{p-1}\Big|\sum_{r=0}^{N-1}\chi_p(dr)\Big|.$$

As in the proof of Lemma 5.2.6 we get

$$\sum_{d=1}^{p-1}\Big|\sum_{r=0}^{N-1}\chi_p(dr)\Big| < p\Big(\log p+\frac{1}{3}\Big),$$

and this concludes the argument. □

Theorem 5.3.25 *Let $p \geq 5$ be a prime number and let $\mathcal{S} = (x_n)_{n=0}^{\infty}$ be a sequence of explicit inversive pseudorandom numbers obtained by* (5.27) *and* (5.28) *with $a, b \in \mathbb{F}_p$ and $a \neq 0$. Then the discrepancy bound*

$$D_N(\mathcal{S}) < N^{-1}\left(2p^{1/2}+1\right)\Big(\log p+\frac{1}{3}\Big)^2+\frac{1}{p}$$

holds for all integers N with $1 \leq N < p$.

Proof. This bound is derived from Lemma 5.3.24 in the same way as Theorem 5.3.13 was derived from Lemma 5.3.12. □

Further results on explicit inversive pseudorandom numbers, including results on the serial test, can be found in the survey article [136]. More recent survey papers containing a lot of relevant information on nonlinear pseudorandom numbers are those of Topuzoğlu and Winterhof [195] and Winterhof [202].

5.4 Pseudorandom bits

So far we have concentrated on random and pseudorandom numbers for Monte Carlo methods and simulation methods. But we should not lose sight of the fact that there are other types of random objects that are consumed on a grand scale, namely random bits. Let us recall, for instance, that encryption by means of stream ciphers is based on the use of sequences of random bits as keystreams (see Section 2.7). We may be tempted to produce the required random bits by coin flips (for example, "head" for 0 and "tail" for 1), but then we run into the same kinds of practical problems as those for physically generated random numbers that we discussed in Subsection 5.1.1. Therefore it is advisable to switch right away to computer-generated random bits called *pseudorandom bits*.

In the utilization of sequences of pseudorandom bits as keystreams, just as in many other applications of pseudorandom bits, we want to avoid any bias between 0 and 1. Thus, the probability of picking 0 should be $\frac{1}{2}$ and the probability of picking 1 should be $\frac{1}{2}$. Furthermore, the choices of bits should be independent in the sense that the current choice of a bit is not influenced by previous choices of bits; think again of fair coin tosses as an illustration. Probability theorists call this a stochastic model for Bernoulli trials, but we will not use this fancy terminology.

Being typical mathematicians, we succumb to the impulse to generalize and we move from the set of bits to an arbitrary finite set S with $b \geq 2$ elements. The generalized stochastic model stipulates now that we pick each element of S with probability $\frac{1}{b}$ and that the choices of elements of S should be independent. We stick to this fair and democratic stochastic model throughout this section. You may think of this model as the discrete analog of the model for uniform pseudorandom numbers employed previously in this chapter. The finite sets S of number-theoretic interest are the least residue system modulo b given by $Z_b = \{0, 1, \ldots, b-1\}$ and the finite field $\mathbb{F}_q$ with q elements in the case where b is a prime power q. The most important special case $b = 2$ corresponding to the set of bits is represented by both Z_2 and $\mathbb{F}_2$.

There is a profusion of plausible properties that we may request for a sequence of pseudorandom elements of S on the basis of the stochastic model above. Many of these properties can be unified into a framework that was popularized by the seminal book of Knuth [81, Section 3.5]. For a sequence $\mathcal{A} = (a_n)_{n=1}^{\infty}$ of elements of S, for integers $k \geq 1$ and $N \geq 1$, and for a k-tuple $\mathbf{s} = (s_0, s_1, \ldots, s_{k-1}) \in S^k$ of elements of S, let $A(\mathbf{s}, N; \mathcal{A})$ denote the number of integers n with $1 \leq n \leq N$ such that the k-tuple $(a_n, a_{n+1}, \ldots, a_{n+k-1})$ of consecutive terms of $\mathcal{A}$ is equal to $\mathbf{s}$. You can picture $A(\mathbf{s}, N; \mathcal{A})$ as follows: slide a window of length k over the sequence $\mathcal{A}$, starting with the window showing $(a_1, a_2, \ldots, a_k)$, and count the number of times you see the k-tuple $\mathbf{s}$ in the window among the first N windows.

Definition 5.4.1 Let S be a finite set with $b \geq 2$ elements and let k be a positive integer. Then a sequence $\mathcal{A}$ of elements of S is *k-distributed in S* if

$$\lim_{N\to\infty} \frac{A(\mathbf{s}, N; \mathcal{A})}{N} = \frac{1}{b^k} \qquad \text{for all } \mathbf{s} \in S^k.$$

A sequence of elements of S is *∞-distributed* (or *completely uniformly distributed*) *in S* if it is k-distributed in S for all integers $k \geq 1$.

Remark 5.4.2 There is an appealing relationship between Definition 5.4.1 and a concept for real numbers, namely that of normality. For an integer $b \geq 2$ and a real number α, let

$$\{\alpha\} = \sum_{n=1}^{\infty} a_n b^{-n}$$

be the unique b-adic expansion of the fractional part $\{\alpha\}$ of α, where $a_n \in Z_b$ for all $n \geq 1$ and $a_n < b-1$ for infinitely many n. Then we can associate with the real number α a unique sequence $\mathcal{A} = (a_n)_{n=1}^{\infty}$ of elements of Z_b, the sequence $\mathcal{A}$ of b-adic digits of $\{\alpha\}$. In the language of Definition 5.4.1, the number α is called *normal to the base b* if the sequence $\mathcal{A}$ is ∞-distributed in Z_b. The theory of normal numbers is a classical and well-studied branch of number theory; see for instance the books [91, Section 1.8] and [151, Chapter 8]. There is an elegant criterion for normality in terms of uniform distribution modulo 1 (see Definition 4.1.8 for the latter notion), according to which the real number α is normal to the base b if and only if the sequence $(b^n\alpha)_{n=1}^{\infty}$ is uniformly distributed modulo 1 (see [91, Chapter 1, Theorem 8.1] and [151, Theorem 8.15]). Another noteworthy result says that "almost all" real numbers are normal to the base b (and in fact simultaneously normal to all bases $b \geq 2$), in the sense that if we pick a real number randomly from the interval $[0, 1)$ equipped with the Lebesgue measure, then with probability 1 this number is normal to the base b (and in fact simultaneously normal to all bases $b \geq 2$); see [91, Chapter 1, Corollaries 8.1 and 8.2] and [151, Theorem 8.11] for different proofs of this result. Thus, in a certain sense, "almost all" sequences of elements of Z_b are ∞-distributed in Z_b.

Example 5.4.3 According to the last part of Remark 5.4.2, there must be a huge variety of sequences of elements of Z_b that are ∞-distributed in Z_b. Nevertheless, it is a nontrivial task to construct such a sequence explicitly. Historically the first construction of an ∞-distributed sequence in Z_b was given by Champernowne [21]. In fact, he constructed a normal number to the base 10, but according to Remark 5.4.2 this is the same as constructing an ∞-distributed sequence in Z_{10}. The sequence $\mathcal{C}$ is obtained by concatenating the digit expansions in base 10 of all positive integers in their natural increasing order. For instance, if you have reached the integer 143, then this yields the three terms $1, 4, 3$ of the sequence $\mathcal{C}$. The beginning of the sequence $\mathcal{C}$ looks like

$$1, 2, 3, 4, 5, 6, 7, 8, 9, 1, 0, 1, 1, 1, 2, 1, 3, 1, 4, \ldots.$$

It is an elementary, but rather boring exercise to show that $\mathcal{C}$ is ∞-distributed in Z_{10}. If you are eager to see the proof, then you are referred to [151, Section 8.4]. More general constructions of this type can be found in [41, Subsection 1.4.4]. It is not known whether interesting numbers such as $\sqrt{2}$, Euler's number e, or π are normal to any base, and such questions belong to the collection of famous open problems in number theory.

We now present the preliminaries of a construction of ∞-distributed sequences in an arbitrary finite field $\mathbb{F}_q$. This construction is of interest since it is based on an important family of periodic sequences with remarkable properties. For an integer $d \geq 1$, we consider the extension field $\mathbb{F}_{q^d}$ of $\mathbb{F}_q$. Let β_d be a primitive element of $\mathbb{F}_{q^d}$ and let $\alpha_d \in \mathbb{F}_{q^d}^*$. We note that the trace map $\mathrm{Tr}_d : \mathbb{F}_{q^d} \to \mathbb{F}_q$ is a surjective linear transformation between the vector spaces $\mathbb{F}_{q^d}$ and $\mathbb{F}_q$ over $\mathbb{F}_q$ (this follows from Theorem 1.4.50). We introduce the sequence $\mathcal{B}_d = (a_n)_{n=1}^{\infty}$ of elements of $\mathbb{F}_q$ by

$$a_n = \mathrm{Tr}_d(\alpha_d \beta_d^n) \qquad \text{for } n = 1, 2, \ldots. \tag{5.29}$$

Since $\beta_d^{q^d-1} = 1$, it is clear that $\mathcal{B}_d$ is a purely periodic sequence with period length $q^d - 1$. The following proposition enunciates impressive equidistribution properties of the sequence $\mathcal{B}_d$ in its full period.

Proposition 5.4.4 *If k and d are integers with $1 \leq k \leq d$ and $\mathbf{s} \in \mathbb{F}_q^k$, then*

$$A(\mathbf{s}, q^d - 1; \mathcal{B}_d) = \begin{cases} q^{d-k} - 1 & \text{if } \mathbf{s} = \mathbf{0}, \\ q^{d-k} & \text{if } \mathbf{s} \neq \mathbf{0}. \end{cases}$$

Proof. We first prove the result for $k = d$. We introduce the linear transformation $L : \mathbb{F}_{q^d} \to \mathbb{F}_q^d$ defined by

$$L(\gamma) = \big(\mathrm{Tr}_d(\alpha_d \gamma), \mathrm{Tr}_d(\alpha_d \beta_d \gamma), \ldots, \mathrm{Tr}_d(\alpha_d \beta_d^{d-1} \gamma)\big) \in \mathbb{F}_q^d \quad \text{for } \gamma \in \mathbb{F}_{q^d}$$

and we claim that L is injective. Thus, let $\gamma \in \mathbb{F}_{q^d}$ be such that $L(\gamma) = \mathbf{0} \in \mathbb{F}_q^d$. Then $\mathrm{Tr}_d(\alpha_d \beta_d^j \gamma) = 0$ for $0 \leq j \leq d-1$. Since $1, \beta_d, \beta_d^2, \ldots, \beta_d^{d-1}$ form a basis of $\mathbb{F}_{q^d}$ over $\mathbb{F}_q$ (see Remark 3.2.7), it follows that $\mathrm{Tr}_d(\alpha_d \gamma \delta) = 0$ for all $\delta \in \mathbb{F}_{q^d}$. This is possible only if $\alpha_d \gamma = 0$ because Tr_d is surjective. Now $\alpha_d \neq 0$, and so $\gamma = 0$, showing that the linear transformation L is indeed injective. Since $\mathbb{F}_{q^d}$ and $\mathbb{F}_q^d$ have the same number of elements, L is even bijective. Note that

$$(a_n, a_{n+1}, \ldots, a_{n+d-1}) = L(\beta_d^n) \qquad \text{for all } n \geq 1$$

by (5.29). Thus $A(\mathbf{s}, q^d - 1; \mathcal{B}_d) = 0$ for $\mathbf{s} = \mathbf{0} \in \mathbb{F}_q^d$. If $\mathbf{s} \in \mathbb{F}_q^d$ with $\mathbf{s} \neq \mathbf{0}$, then there exists a unique $\gamma \in \mathbb{F}_{q^d}^*$ with $L(\gamma) = \mathbf{s}$, and so a unique n with $1 \leq n \leq q^d - 1$ such that $L(\beta_d^n) = \mathbf{s}$. This proves that $A(\mathbf{s}, q^d - 1; \mathcal{B}_d) = 1$.

Now we examine the case where $1 \leq k < d$. If $\pi(\mathbf{v}) \in \mathbb{F}_q^k$ denotes the projection of $\mathbf{v} \in \mathbb{F}_q^d$ onto its first k coordinates, then for every $\mathbf{s} \in \mathbb{F}_q^k$ we get the formula

$$A(\mathbf{s}, q^d - 1; \mathcal{B}_d) = \sum_{\substack{\mathbf{v} \in \mathbb{F}_q^d \\ \pi(\mathbf{v}) = \mathbf{s}}} A(\mathbf{v}, q^d - 1; \mathcal{B}_d).$$

The proof is completed by using the result that we have already shown for the case $k = d$. □

We need also a result on the counting function $A(\mathbf{s}, N; \mathcal{B}_d)$ for parts of the period, that is, for $1 \leq N < q^d - 1$. We can take recourse to earlier methods in this chapter in order to obtain such a result.

Proposition 5.4.5 *If k and d are integers with $1 \leq k \leq d$ and $\mathbf{s} \in \mathbb{F}_q^k$, then*

$$|A(\mathbf{s}, N; \mathcal{B}_d) - Nq^{-k}| < q^{d/2}(\log q^d + 1) \qquad \textit{for } 1 \leq N < q^d - 1.$$

Proof. This case can arise only if $q^d \geq 3$, and so we assume this inequality. We fix an integer N with $1 \leq N < q^d - 1$, a nontrivial additive character χ of $\mathbb{F}_q$, and $\mathbf{s} = (s_0, s_1, \ldots, s_{k-1}) \in \mathbb{F}_q^k$. We put

$$\mathbf{a}_n = (a_n, a_{n+1}, \ldots, a_{n+k-1}) \in \mathbb{F}_q^k \qquad \text{for } n = 1, 2, \ldots,$$

where $a_1, a_2, \ldots$ are the terms of the sequence $\mathcal{B}_d$. Then, using the dot product on $\mathbb{F}_q^k$, we can write

$$\begin{aligned} A(\mathbf{s}, N; \mathcal{B}_d) &= \sum_{n=1}^{N} \prod_{j=0}^{k-1} \Big(\frac{1}{q} \sum_{c \in \mathbb{F}_q} \chi(c(a_{n+j} - s_j)) \Big) \\ &= q^{-k} \sum_{n=1}^{N} \sum_{\mathbf{c} \in \mathbb{F}_q^k} \chi(\mathbf{c} \cdot \mathbf{a}_n - \mathbf{c} \cdot \mathbf{s}) \\ &= q^{-k} \sum_{\mathbf{c} \in \mathbb{F}_q^k} \chi(-\mathbf{c} \cdot \mathbf{s}) \sum_{n=1}^{N} \chi(\mathbf{c} \cdot \mathbf{a}_n). \end{aligned}$$

By splitting off the contribution from $\mathbf{c} = \mathbf{0} \in \mathbb{F}_q^k$ and using the triangle inequality, we obtain

$$|A(\mathbf{s}, N; \mathcal{B}_d) - Nq^{-k}| \leq q^{-k} \sum_{\mathbf{c} \in \mathbb{F}_q^k \setminus \{\mathbf{0}\}} \Big| \sum_{n=1}^{N} \chi(\mathbf{c} \cdot \mathbf{a}_n) \Big|. \qquad (5.30)$$

For $\mathbf{c} = (c_0, c_1, \dots, c_{k-1}) \in \mathbb{F}_q^k$ with $\mathbf{c} \neq \mathbf{0} \in \mathbb{F}_q^k$, we use (5.29) to get for all $n \geq 1$,

$$\begin{aligned}\mathbf{c} \cdot \mathbf{a}_n &= c_0 a_n + c_1 a_{n+1} + \cdots + c_{k-1} a_{n+k-1} \\ &= \mathrm{Tr}_d(\alpha_d(c_0 + c_1 \beta_d + \cdots + c_{k-1}\beta_d^{k-1})\beta_d^n) = \mathrm{Tr}_d(\gamma \beta_d^n)\end{aligned}$$

with some $\gamma \in \mathbb{F}_{q^d}$. It is important to observe that $\gamma \neq 0$ since $k - 1 < d$. Now $\sigma(\delta) = \chi(\mathrm{Tr}_d(\delta))$ for all $\delta \in \mathbb{F}_{q^d}$ defines a nontrivial additive character σ of $\mathbb{F}_{q^d}$, and so we can write

$$\sum_{n=1}^{N} \chi(\mathbf{c} \cdot \mathbf{a}_n) = \sum_{n=1}^{N} \sigma(\gamma \beta_d^n). \tag{5.31}$$

The character sum on the right-hand side of (5.31) is treated by the method in the proof of Lemma 5.2.6. For $T = q^d - 1$ we put $\chi_T(z) = e^{2\pi i z/T}$ for all $z \in \mathbb{Z}$ as in the proof of that lemma. Then we arrive at the bound

$$\Big|\sum_{n=1}^{N} \sigma(\gamma\beta_d^n)\Big| \leq \frac{1}{T} \sum_{h=0}^{T-1} \Big|\sum_{r=0}^{N-1} \chi_T(hr)\Big| \Big|\sum_{n=0}^{T-1} \sigma(\gamma\beta_d^n)\chi_T(hn)\Big|.$$

We consider the contribution from $h = 0$, use (1.9), and obtain

$$\sum_{n=0}^{T-1} \sigma(\gamma\beta_d^n) = \sum_{\delta \in \mathbb{F}_{q^d}^*} \sigma(\delta) = \sum_{\delta \in \mathbb{F}_{q^d}} \sigma(\delta) - 1 = -1. \tag{5.32}$$

Noting also that $N < T$, we get

$$\Big|\sum_{n=1}^{N} \sigma(\gamma\beta_d^n)\Big| < 1 + \frac{1}{T} \sum_{h=1}^{T-1} \Big|\sum_{r=0}^{N-1} \chi_T(hr)\Big| \Big|\sum_{n=0}^{T-1} \sigma(\gamma\beta_d^n)\chi_T(hn)\Big|. \tag{5.33}$$

For the last sum, we proceed as in the proof of Lemma 5.2.4, with $\mathbb{F}_p$ replaced by $\mathbb{F}_{q^d}$ and χ_p replaced by σ. This yields

$$\Big|\sum_{n=0}^{T-1} \sigma(\gamma\beta_d^n)\chi_T(hn)\Big| \leq q^{d/2} \qquad \text{for } 1 \leq h \leq T-1.$$

By using this inequality in (5.33), we obtain

$$\Big|\sum_{n=1}^{N} \sigma(\gamma\beta_d^n)\Big| < 1 + \frac{q^{d/2}}{T} \sum_{h=1}^{T-1} \Big|\sum_{r=0}^{N-1} \chi_T(hr)\Big|.$$

The sum over h was bounded in the proof of Lemma 5.2.6, and thus we get

$$\Big|\sum_{n=1}^{N} \sigma(\gamma\beta_d^n)\Big| < 1 + q^{d/2}\Big(\log T + \frac{1}{3}\Big) < q^{d/2}(\log q^d + 1).$$

Now we combine this inequality with (5.30) and (5.31) and we arrive at the desired result. □

Remark 5.4.6 The terms a_n in (5.29) of the sequence $\mathcal{B}_d$ can be computed efficiently. Let $f \in \mathbb{F}_q[x]$ be the minimal polynomial of β_d over $\mathbb{F}_q$, which is thus a primitive polynomial over $\mathbb{F}_q$ of degree d. We write $f(x) = x^d - \sum_{j=0}^{d-1} e_j x^j$ with $e_j \in \mathbb{F}_q$ for $0 \le j \le d-1$. Then

$$a_{n+d} - \sum_{j=0}^{d-1} e_j a_{n+j} = \mathrm{Tr}_d\Big(\alpha_d \beta_d^n \Big(\beta_d^d - \sum_{j=0}^{d-1} e_j \beta_d^j\Big)\Big) = \mathrm{Tr}_d(\alpha_d \beta_d^n f(\beta_d)) = \mathrm{Tr}_d(0) = 0$$

for all $n \ge 1$, and so the a_n satisfy the linear recurrence relation over $\mathbb{F}_q$ (of order d)

$$a_{n+d} = \sum_{j=0}^{d-1} e_j a_{n+j} \qquad \text{for } n = 1, 2, \ldots.$$

Once the initial values $a_1, \ldots, a_d$ have been computed, the remaining terms of the sequence $\mathcal{B}_d$ can be quickly generated by this linear recurrence relation. Proposition 5.4.4 shows that the d-tuples $(a_n, a_{n+1}, \ldots, a_{n+d-1})$, $n = 1, \ldots, q^d - 1$, run exactly through all nonzero d-tuples in $\mathbb{F}_q^d$. This implies that $q^d - 1$ is the least period length of $\mathcal{B}_d$. Any linear recurrence relation over $\mathbb{F}_q$ of order d generates a periodic sequence with least period length at most $q^d - 1$, and for this reason the sequence $\mathcal{B}_d$ is called a *maximal period sequence* over $\mathbb{F}_q$.

Now we are ready to describe the construction of ∞-distributed sequences in an arbitrary finite field $\mathbb{F}_q$, following the paper [138]. For each positive integer d, let $\mathcal{B}_d$ be the maximal period sequence over $\mathbb{F}_q$ constructed above and let $\mathcal{T}_d$ be the block (or the initial segment) consisting of the first $q^d - 1$ terms of $\mathcal{B}_d$, that is, $\mathcal{T}_d$ is the first full period of $\mathcal{B}_d$. By concatenating $\mathcal{T}_1, \mathcal{T}_2, \ldots$, we get the sequence $\mathcal{B}$ of elements of $\mathbb{F}_q$.

Theorem 5.4.7 *The sequence $\mathcal{B}$ obtained by concatenating the blocks $\mathcal{T}_1, \mathcal{T}_2, \ldots$ is ∞-distributed in $\mathbb{F}_q$.*

Proof. We put $M_h = \sum_{d=1}^{h} (q^d - 1)$ for integers $h \ge 1$, so that M_h is the total number of terms after the concatenation of $\mathcal{T}_1, \ldots, \mathcal{T}_h$. We select an integer $k \ge 1$ and we want to prove that $\mathcal{B}$ is k-distributed in $\mathbb{F}_q$. Choose $\mathbf{s} \in \mathbb{F}_q^k$ and an integer $N > M_k$. Then there exists a unique integer $r \ge k + 1$ with $M_{r-1} < N \le M_r$. The first N terms of $\mathcal{B}$ consist therefore of $\mathcal{T}_1, \ldots, \mathcal{T}_{r-1}$ and the first $N - M_{r-1}$ terms of $\mathcal{T}_r$. Hence

$$A(\mathbf{s}, N; \mathcal{B}) = \sum_{d=1}^{r-1} A(\mathbf{s}, q^d - 1; \mathcal{B}_d) + A(\mathbf{s}, N - M_{r-1}; \mathcal{B}_r) + O(kr).$$

The correction term $O(kr)$, where here and in the rest of the proof all implied constants are absolute, reflects the possible errors in the counts at the interfaces between $\mathcal{T}_d$ and $\mathcal{T}_{d+1}$ for $d = 1, \dots, r$. For $d \geq k$ we can apply Proposition 5.4.4, which we write in the form

$$A(\mathbf{s}, q^d - 1; \mathcal{B}_d) = (q^d - 1)q^{-k} + O(1).$$

For $1 \leq d < k$ it is trivial that

$$A(\mathbf{s}, q^d - 1; \mathcal{B}_d) = (q^d - 1)q^{-k} + O(q^d).$$

Therefore

$$A(\mathbf{s}, N; \mathcal{B}) = M_{r-1}q^{-k} + A(\mathbf{s}, N - M_{r-1}; \mathcal{B}_r) + O(kr + q^k).$$

Next an application of Proposition 5.4.5 yields

$$A(\mathbf{s}, N - M_{r-1}; \mathcal{B}_r) = (N - M_{r-1})q^{-k} + O(q^{r/2} \log q^r).$$

It follows that

$$A(\mathbf{s}, N; \mathcal{B}) = Nq^{-k} + O(kr + q^k + rq^{r/2} \log q).$$

Now $N \geq M_{r-1} + 1 \geq q^{r-1}$, and so

$$\frac{A(\mathbf{s}, N; \mathcal{B})}{N} = q^{-k} + O(krq^{1-r} + q^{k+1-r} + rq^{1-r/2} \log q).$$

If we now let $N \to \infty$, then $r \to \infty$, and we infer that $\mathcal{B}$ is indeed k-distributed in $\mathbb{F}_q$. Since $k \in \mathbb{N}$ is arbitrary, it follows that $\mathcal{B}$ is ∞-distributed in $\mathbb{F}_q$. □

For a purely periodic sequence $\mathcal{A} = (a_n)_{n=1}^{\infty}$ of elements of $\mathbb{F}_q$ with least period length T, it is customary to consider the *correlation coefficient*

$$C_h(\mathcal{A}) = \sum_{n=1}^{T} \chi(a_n - a_{n+h}),$$

where χ is a fixed nontrivial additive character of $\mathbb{F}_q$ and h is a positive integer. Correlation coefficients are special instances of autocorrelation functions which will be introduced in Definition 6.4.15. For a good sequence $\mathcal{A}$ of pseudorandom elements of $\mathbb{F}_q$, the absolute value of the correlation coefficient $C_h(\mathcal{A})$ should be small compared to T for many values of h. The maximal period sequences $\mathcal{B}_d$ constructed above are particularly well behaved in this respect.

Theorem 5.4.8 *For a maximal period sequence $\mathcal{B}_d$ over $\mathbb{F}_q$ with least period length $T = q^d - 1$, its correlation coefficients are given by*

$$C_h(\mathcal{B}_d) = \begin{cases} T & \text{if } h \equiv 0 \pmod{T}, \\ -1 & \text{otherwise.} \end{cases}$$

Proof. The case $h \equiv 0 \pmod{T}$ is trivial since $\mathcal{B}_d$ has period length T. If $h \not\equiv 0 \pmod{T}$, then (5.29) yields

$$C_h(\mathcal{B}_d) = \sum_{n=1}^{T} \chi(\mathrm{Tr}_d\left(\alpha_d \beta_d^n) - \mathrm{Tr}_d(\alpha_d \beta_d^{n+h})\right) = \sum_{n=1}^{T} \chi\left(\mathrm{Tr}_d(\alpha_d(1-\beta_d^h)\beta_d^n)\right).$$

Now $\beta_d^h \neq 1$, and so $\gamma := \alpha_d(1-\beta_d^h) \in \mathbb{F}_{q^d}^*$. With σ as in the proof of Proposition 5.4.5, we obtain

$$C_h(\mathcal{B}_d) = \sum_{n=1}^{T} \sigma(\gamma \beta_d^n) = -1$$

by (5.32). □

Maximal period sequences over the binary field $\mathbb{F}_2$ are the building blocks for keystreams that are used in practice in stream ciphers. Since maximal period sequences can be generated by linear recurrence relations (see Remark 5.4.6), their structure is too simple for keystreams, and so some features of nonlinearity have to be introduced. A common procedure is to combine several maximal period sequences over $\mathbb{F}_2$ by a nonlinear combining function. So if $m \geq 2$ maximal period sequences over $\mathbb{F}_2$ are combined, the combining function is a nonlinear function $g : \mathbb{F}_2^m \to \mathbb{F}_2$. For $n = 1, 2, \ldots$, the nth term of the keystream is $g(a_n^{(1)}, \ldots, a_n^{(m)})$, where $a_n^{(j)}$ is the nth term of the jth maximal period sequence for $1 \leq j \leq m$. A discussion of the choice of combining functions can be found in [116, Section 6.3].

Keystreams for stream ciphers should have good statistical properties, for instance in the sense of being k-distributed in $\mathbb{F}_2$ for large values of k. But they should also have properties that can be roughly described by saying that the keystream is patternless and unpredictable, so that attackers cannot figure out the algorithm (or crucial parameters in the algorithm) by which the keystream is generated. These types of properties are analyzed by complexity theory, which is a big and fundamental branch of theoretical computer science. In this area, various complexity measures have been devised in order to assess how close to random a sequence of bits is. The general idea is to measure the level of complexity of the simplest algorithm (or equivalently of the simplest machine) that can generate the given sequence of bits. The concrete complexity measure depends on which family of algorithms (or machines) one allows in the competition. The most ambitious approach considers *all* machines that are relevant for computer science, namely all (self-delimiting) Turing machines, and this leads to the concept of the (self-delimiting) *Kolmogorov complexity* of a sequence of bits. Intuitively, you may think of the Kolmogorov complexity of a sequence of bits as the length of the shortest computer program for generating the sequence or an initial segment thereof. A rich and beautiful theory of the Kolmogorov complexity and of its relationship with randomness properties of sequences of bits was developed by computer scientists (see [101] for a survey). The only hitch in this theory is that it can be proved that the Kolmogorov complexity is in general not efficiently computable, and this is of course a severe blow to the practical utility of this complexity measure.

At the other end of the hierarchy of complexity measures is one where the only algorithms we allow are linear recurrence relations. This leads to the concept of the *linear complexity* of a periodic sequence of bits, which is simply the least order of a linear recurrence relation that generates the sequence (compare also with Section 2.7). The linear complexity of an initial segment of an arbitrary sequence of bits is defined analogously. The linear complexity has the great advantage over the Kolmogorov complexity that it can be computed by a polynomial-time algorithm, which happens to be the Berlekamp-Massey algorithm originally designed for coding theory (see Section 3.6). Surveys of the linear complexity and of related complexity measures can be found in the articles [114] and [201]. The monograph [32] is devoted to the complexity analysis of sequences generated by number-theoretic methods.

5.5 A glimpse of advanced topics

One of the disadvantages of sequences of linear congruential pseudorandom numbers is that their least period length cannot exceed the modulus. We can overcome this drawback by replacing the first-order linear recurrence relation (5.2) by a linear recurrence relation of higher order $r \geq 2$, thus arriving at the *multiple-recursive method*. We choose a large prime number p as the modulus and coefficients $c_0, c_1, \ldots, c_{r-1} \in Z_p = \{0, 1, \ldots, p-1\}$. Then we generate a sequence $(z_n)_{n=0}^{\infty}$ of elements of Z_p by the linear recurrence relation

$$z_{n+r} \equiv \sum_{j=0}^{r-1} c_j z_{n+j} \pmod{p} \qquad \text{for } n = 0, 1, \ldots. \tag{5.34}$$

It is assumed that not all initial values $z_0, z_1, \ldots, z_{r-1}$ are 0. In analogy with (5.3), a sequence $(x_n)_{n=0}^{\infty}$ of *multiple-recursive pseudorandom numbers* is obtained by $x_n = z_n/p \in [0, 1)$ for $n = 0, 1, \ldots$. In order to maximize the least period length of this sequence, we consider the so-called characteristic polynomial $f(x) = x^r - \sum_{j=0}^{r-1} c_j x^j$ of the linear recurrence relation as a polynomial over the finite field $\mathbb{F}_p$ and we suppose that f is a primitive polynomial over $\mathbb{F}_p$. Then $\mathrm{per}(x_n) = \mathrm{per}(z_n) = p^r - 1$ and the sequence $(z_n)_{n=0}^{\infty}$ is a maximal period sequence over $\mathbb{F}_p$ in the sense of Remark 5.4.6. Therefore the excellent equidistribution properties of maximal period sequences established in Propositions 5.4.4 and 5.4.5 apply to the sequence $(z_n)_{n=0}^{\infty}$ for dimensions $k \leq r$.

Because of the linearity of the recurrence relation generating the sequence $(z_n)_{n=0}^{\infty}$, multiple-recursive pseudorandom numbers still show a lattice structure or grid structure, just like linear congruential pseudorandom numbers (compare with Subsection

5.2.2). There are tools such as the so-called spectral test that allow us to discriminate between good and bad parameters for multiple-recursive pseudorandom numbers, and the spectral test can also be applied to pick out good multipliers in the linear congruential method. An excellent account of the spectral test and of related structural and statistical tests is given in the survey article [95].

What you gain on the roundabouts, you lose on the swings, and here for multiple-recursive pseudorandom numbers we win handsomely with the least period length, but we pay a price in terms of poor discretization and discrepancy. For $\mathrm{per}(x_n) = p^r - 1$ as above, we would expect a discretization roughly of size p^{-r}, but the x_n are rational numbers with denominator p and therefore yield a discretization of size p^{-1}. Consequently, the star discrepancy of any initial segment of the sequence $(x_n)_{n=0}^{\infty}$ is at least p^{-1}, which in most cases is too big to be anywhere near the law of the iterated logarithm for the star discrepancy.

A smart way to go is to combine large least period length with fine discretization, and this is carried out in the *digital multistep method*. Here we choose $p = 2$ and a large order r of the linear recurrence relation (5.34). We use (5.34) to generate a maximal period sequence $(z_n)_{n=0}^{\infty}$ over $\mathbb{F}_2$ with $\mathrm{per}(z_n) = 2^r - 1$. Note that it is no problem to achieve huge least period lengths like $2^{1000} - 1$ or $2^{5000} - 1$ since r can be chosen independently of the available processor and only the extremely fast binary arithmetic is needed for generating the z_n. In contrast, the choice of practical moduli and therefore of least period lengths in the linear congruential method is limited by the word size of the processor (compare with the discussion in Subsection 5.2.1). But how do we produce pseudorandom numbers in $[0, 1)$ from the sequence $(z_n)_{n=0}^{\infty}$ of bits? Well, we choose an integer k with $2 \le k \le r$ and $\gcd(k, 2^r - 1) = 1$ and we put

$$x_n = \sum_{i=1}^{k} z_{kn+i-1} 2^{-i} \in [0, 1) \qquad \text{for } n = 0, 1, \dots. \tag{5.35}$$

In words, the numbers x_n are obtained by splitting up the sequence $(z_n)_{n=0}^{\infty}$ into contiguous blocks of length k and then interpreting each block as the dyadic expansion of a number in $[0, 1)$. The numbers x_n defined by (5.35) are called *digital multistep pseudorandom numbers*. The condition $\gcd(k, 2^r - 1) = 1$ guarantees that $\mathrm{per}(x_n) = 2^r - 1$. The discretization is 2^{-k} and we can choose very large values for k. Almost perfect equidistribution holds for dimensions $s \le r/k$ (see [134, Theorem 9.2]).

There is an astonishing connection between digital multistep pseudorandom numbers and the theory of digital nets presented in Subsection 4.4.2. For a dimension $s > r/k$, we introduce the points

$$\mathbf{x}_n = (x_n, x_{n+1}, \dots, x_{n+s-1}) \in [0, 1)^s \qquad \text{for } n = 0, 1, \dots,$$

where the sequence $(x_n)_{n=0}^{\infty}$ is given by (5.35). Then it can be verified that the 2^r points $\mathbf{0}, \mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{2^r-2}$ form a digital (t, r, s)-net over $\mathbb{F}_2$ with a quality parameter

t that depends in a known way on r, k, and the characteristic polynomial f of the linear recurrence relation generating the sequence $(z_n)_{n=0}^{\infty}$ (see [130] and [134, Theorem 9.5]). There are results guaranteeing that, by an appropriate choice of the characteristic polynomial f, the quality parameter t of the digital net and the star discrepancy of the points $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{2^r-2}$ can be made small (see [134, Theorems 9.7 and 9.8]).

The basic idea of the digital multistep method, namely to create (pseudo)randomness by combining finite-field arithmetic and real arithmetic via (5.34) and (5.35), is exploited also in the *digital inversive method* due to Eichenauer-Herrmann and Niederreiter [46]. We select an integer $k \geq 1$ which serves as the precision (like $k = 32$ or $k = 64$) and we consider the finite field $\mathbb{F}_q$ with $q = 2^k$ elements. For all $\gamma \in \mathbb{F}_q$, we use the notation $\overline{\gamma} = \gamma^{-1} \in \mathbb{F}_q$ if $\gamma \neq 0$ and $\overline{\gamma} = 0 \in \mathbb{F}_q$ if $\gamma = 0$. Now we proceed in analogy with (5.19), but we work in the finite field $\mathbb{F}_q$ rather than in $\mathbb{F}_p$. Concretely, we choose parameters $\alpha, \beta \in \mathbb{F}_q^*$ and an initial value $\gamma_0 \in \mathbb{F}_q$, and then we generate the sequence $(\gamma_n)_{n=0}^{\infty}$ of elements of $\mathbb{F}_q$ by the recursion

$$\gamma_{n+1} = \alpha\overline{\gamma_n} + \beta \qquad \text{for } n = 0, 1, \dots .$$

Next, for $n = 0, 1, \dots$, let $(y_n^{(1)}, \dots, y_n^{(k)}) \in \mathbb{F}_2^k$ be the coordinate vector of γ_n relative to a fixed ordered basis of $\mathbb{F}_q$ over $\mathbb{F}_2$. Finally, we identify $\mathbb{F}_2$ with $Z_2 = \{0, 1\}$ and we introduce a sequence $(x_n)_{n=0}^{\infty}$ of *digital inversive pseudorandom numbers* by

$$x_n = \sum_{i=1}^{k} y_n^{(i)} 2^{-i} \in [0, 1) \qquad \text{for } n = 0, 1, \dots .$$

Under similar conditions as in Theorem 5.3.16, we get $\mathrm{per}(x_n) = q = 2^k$. Digital inversive pseudorandom numbers possess agreeable properties with regard to the uniformity test and the serial test (see [46] and [143]). Furthermore, these pseudorandom numbers allow an efficient implementation. For general finite fields $\mathbb{F}_q$ (including finite prime fields $\mathbb{F}_p$), the computation of the multiplicative inverse in $\mathbb{F}_q$ requires $O(\log q)$ multiplications in $\mathbb{F}_q$; simply note that $\gamma^{-1} = \gamma^{q-2}$ for all $\gamma \in \mathbb{F}_q^*$ and apply the square-and-multiply algorithm in Algorithm 2.3.9. But in the special case $q = 2^k$, a clever algorithm due to Itoh and Tsujii [72] permits the computation of the multiplicative inverse in $\mathbb{F}_q$ with $O(\log \log q)$ multiplications in $\mathbb{F}_q$, and this is of course an enormous speedup. For values of q of practical interest such as $q = 2^{32}$, the Itoh-Tsujii algorithm computes the multiplicative inverse in $\mathbb{F}_q$ basically in constant time.

Parallelized Monte Carlo methods and simulation methods employ sequences of pseudorandom vectors. The analog of the linear congruential method in this context is the *matrix method*. For a given dimension $k \geq 2$, we choose a large prime number p and a nonsingular $k \times k$ matrix A over the finite field $\mathbb{F}_p$. Then we generate a sequence $\mathbf{z}_0, \mathbf{z}_1, \dots$ of row vectors in $\mathbb{F}_p^k$ by starting from an initial vector $\mathbf{z}_0 \neq \mathbf{0} \in \mathbb{F}_p^k$

and using the recursion

$$\mathbf{z}_{n+1} = \mathbf{z}_n A \qquad \text{for } n = 0, 1, \ldots.$$

Now we identify $\mathbb{F}_p$ with $Z_p = \{0, 1, \ldots, p-1\}$ and we derive the sequence of pseudorandom vectors

$$\mathbf{x}_n = \frac{1}{p}\mathbf{z}_n \in [0,1)^k \qquad \text{for } n = 0, 1, \ldots.$$

It is obvious that the sequence $(\mathbf{x}_n)_{n=0}^{\infty}$ is purely periodic with $\mathrm{per}(\mathbf{x}_n) \le p^k - 1$. We get $\mathrm{per}(\mathbf{x}_n) = p^k - 1$ if and only if the characteristic polynomial of the matrix A is primitive over $\mathbb{F}_p$ (see [134, Theorem 10.2]).

From a sequence $\mathbf{z}_0, \mathbf{z}_1, \ldots$ of pseudorandom vectors in $\mathbb{F}_p^k$ we can also obtain a sequence of pseudorandom numbers in $[0,1)$. To this end, we choose $p = 2$ and let the integer k be a precision, say $k = 32$ or $k = 64$. We write

$$\mathbf{z}_n = (z_n^{(1)}, \ldots, z_n^{(k)}) \in \mathbb{F}_2^k \qquad \text{for } n = 0, 1, \ldots$$

and then we produce the pseudorandom numbers

$$x_n = \sum_{i=1}^{k} z_n^{(i)} 2^{-i} \in [0,1) \qquad \text{for } n = 0, 1, \ldots.$$

If we use a sophisticated method for pseudorandom vector generation such as the so-called multiple-recursive matrix method (the vector analog of the multiple-recursive method), then we obtain a sequence $(x_n)_{n=0}^{\infty}$ of pseudorandom numbers with many desirable properties (see [137]). A special instance of this approach yields the famous *Mersenne twister* invented by Matsumoto and Nishimura [111]. The Mersenne twister is a marvel of design: it produces periodic sequences of pseudorandom numbers with the least period length being the huge Mersenne prime $2^{19937} - 1$ and with almost perfect equidistribution properties all the way up to the dimension 623. The sequences of pseudorandom numbers generated by the Mersenne twister pass numerous statistical tests for randomness and they are now widely used in practice. You can find some information on Mersenne primes in Subsection 6.6.3.

Exercises

5.1 For an integer $m \ge 3$, generate a sequence $z_0, z_1, \ldots$ of elements of Z_m by $z_{n+2} \equiv z_{n+1} + z_n \pmod{m}$ for $n = 0, 1, \ldots$ with arbitrary initial values z_0 and z_1. Derive a sequence $(x_n)_{n=0}^{\infty}$ of pseudorandom numbers by putting $x_n = z_n/m$ for all $n \ge 0$. Show that the sequence $(x_n)_{n=0}^{\infty}$ badly fails the three-dimensional permutation test, in the sense that the ordering $x_n < x_{n+2} < x_{n+1}$ never occurs in this sequence.

5.2 For an integer $m \geq 2$, let $(a_n)_{n=0}^{\infty}$ and $(b_n)_{n=0}^{\infty}$ be purely periodic sequences of elements of Z_m such that $\operatorname{per}(a_n)$ and $\operatorname{per}(b_n)$ are coprime. Prove that the sequence $(z_n)_{n=0}^{\infty}$ of elements of Z_m defined by $z_n \equiv a_n + b_n \pmod{m}$ for all $n \geq 0$ satisfies $\operatorname{per}(z_n) = \operatorname{per}(a_n)\operatorname{per}(b_n)$.

5.3 Let $m = \prod_{j=1}^{k} p_j^{e_j}$ be the canonical factorization of the integer $m \geq 2$ and let $a, z_0 \in \mathbb{Z}$ with $\gcd(a, m) = \gcd(z_0, m) = 1$. Prove that the least period length of $(a^n z_0)_{n=0}^{\infty}$ considered as a sequence modulo m is $\operatorname{lcm}(T_1, \ldots, T_k)$, where T_j for $1 \leq j \leq k$ is the least period length of $(a^n z_0)_{n=0}^{\infty}$ considered as a sequence modulo $p_j^{e_j}$.

5.4 For every integer $m \geq 2$ and every map $\psi : Z_m \to Z_m$, prove that any sequence $z_0, z_1, \ldots$ of elements of Z_m generated by $z_{n+1} = \psi(z_n)$ for $n = 0, 1, \ldots$ with an arbitrary initial value z_0 is ultimately periodic. Provide a reasonable sufficient condition for the sequence to be purely periodic.

5.5 Consider the inhomogeneous case of the linear congruential method in Remark 5.2.3 for $m = p^k$ with a prime number p and an integer $k \geq 1$. Assume also that $a \not\equiv 1 \pmod{p}$ and that $(a-1)z_0 + c \neq 0$. Let r be the largest integer such that p^r divides $(a-1)z_0 + c$ and suppose that $k \geq r$. Prove that $\operatorname{per}(z_n)$ is equal to the multiplicative order of a modulo p^{k-r}.

5.6 Consider the linear congruential method in Remark 5.2.3 for $m = p^k$ with a prime number p and an integer $k \geq 2$, where we allow also the case $c = 0$. Let T, respectively T_1, be the least period length of the sequence $(z_n)_{n=0}^{\infty}$ considered as a sequence modulo m, respectively modulo p^{k-1}, and suppose that $T = pT_1$. Prove that

$$\sum_{n=0}^{T-1} \chi_m(bz_n) = 0 \qquad \text{for all integers } b \not\equiv 0 \pmod{p}.$$

5.7 Let χ be a nontrivial additive character of the finite field $\mathbb{F}_q$. Let $G(\psi, \chi)$ be the Gauss sum defined in Exercise 1.34 and put $G(\psi_0, \chi) = -1$ for the trivial multiplicative character ψ_0 of $\mathbb{F}_q$. Prove that

$$\chi(c) = \frac{1}{q-1} \sum_{\psi} G(\psi, \chi)\overline{\psi(c)} \qquad \text{for all } c \in \mathbb{F}_q^*,$$

where the sum is extended over all multiplicative characters ψ of $\mathbb{F}_q$ and where the bar denotes complex conjugation.

5.8 Prove the following version of Lemma 5.2.4 for an arbitrary finite field $\mathbb{F}_q$. Let χ be a nontrivial additive character of $\mathbb{F}_q$, let $a, b \in \mathbb{F}_q^*$, and let T be the order of a in the multiplicative group $\mathbb{F}_q^*$. Then

$$\Big|\sum_{n=0}^{T-1} \chi(ba^n)\Big| \leq q^{1/2} - T(q^{1/2}+1)^{-1}.$$

(Hint: use the preceding exercise and Exercise 1.34.)

5.9 Let p be an odd prime number, let $a \in \mathbb{Z}$ with $\gcd(a,p)=1$, and assume that the multiplicative order T of a modulo p satisfies $T \geq N := (p-1)/2$. Prove that there exists an integer b with $\gcd(b,p)=1$ such that

$$\Big|\sum_{n=0}^{N-1} \chi_p(ba^n)\Big| \geq \frac{1}{2}(p+1)^{1/2}.$$

This shows that Lemma 5.2.6 is in general best possible up to the logarithmic factor. (Hint: consider $\sum_{b=1}^{p-1} \big|\sum_{n=0}^{N-1} \chi_p(ba^n)\big|^2$.)

5.10 Let p, a, T, and N be as in the preceding exercise. Prove that there exists an integer b with $\gcd(b,p)=1$ such that the star discrepancy D_N^* of the point set consisting of the fractional parts $\{a^n b/p\}$ with $n = 0, 1, \ldots, N-1$ satisfies

$$D_N^* \geq \frac{(p+1)^{1/2}}{8N}.$$

This shows that Theorem 5.2.7 is in general best possible up to the logarithmic factors.

5.11 Let $m = 2^k$ with an integer $k \geq 1$ and generate a sequence $z_0, z_1, \ldots$ of elements of Z_m by $z_{n+2} \equiv z_{n+1} + z_n \pmod{m}$ for $n = 0, 1, \ldots$ with initial values z_0 and z_1 that are not both even. Prove that $\mathrm{per}(z_n) = 3 \cdot 2^{k-1}$.

5.12 Let p be a prime number, let $s \in \mathbb{N}$ with $s < p$, and let $g(x) \in \mathbb{F}_p[x]$ with $s \leq \deg(g(x)) < p$. Prove that for every nonzero vector $(h_0, h_1, \ldots, h_{s-1}) \in \mathbb{F}_p^s$, the polynomial $\sum_{j=0}^{s-1} h_j g(x+j) \in \mathbb{F}_p[x]$ has positive degree.

5.13 Let $p \geq 5$ be a prime number and let $(x_n)_{n=0}^{\infty}$ be a sequence of explicit nonlinear pseudorandom numbers generated by (5.15) and (5.18) with $3 \leq \deg(g) \leq p-2$. For a dimension $s \leq \deg(g)$, let $\mathcal{P}_p^{(s)}$ be the point set consisting of the points

$$\mathbf{x}_n = (x_n, x_{n+1}, \ldots, x_{n+s-1}) \in [0,1)^s \qquad \text{for } n = 0, 1, \ldots, p-1.$$

Prove that the discrepancy $D_p(\mathcal{P}_p^{(s)})$ of $\mathcal{P}_p^{(s)}$ satisfies

$$D_p(\mathcal{P}_p^{(s)}) = O\left(p^{-1/2}(\log p)^s\right)$$

with an implied constant depending only on $\deg(g)$. (Hint: use Proposition 4.3.1 and Exercise 5.12.)

5.14 Let the sequence $(x_n)_{n=0}^{\infty}$ be as in the preceding exercise. For a dimension $s \leq \deg(g) - 1$ and for an integer N with $1 \leq N < p$, let $\mathcal{P}_N^{(s)}$ be the point set consisting of the points

$$\mathbf{x}_n = (x_n, x_{n+1}, \ldots, x_{n+s-1}) \in [0,1)^s \qquad \text{for } n = 0, 1, \ldots, N-1.$$

Prove that the discrepancy $D_N(\mathcal{P}_N^{(s)})$ of $\mathcal{P}_N^{(s)}$ satisfies

$$D_N(\mathcal{P}_N^{(s)}) = O\left(N^{-1} p^{1/2} (\log p)^{s+1}\right)$$

with an implied constant depending only on $\deg(g)$.

5.15 Let p be a prime number and let $g(x) \in \mathbb{F}_p[x]$ with $0 \leq \deg(g(x)) = d < p$. Prove that there exists an element $c \in \mathbb{F}_p$ such that

$$\sum_{j=0}^{d} (-1)^{d-j} \binom{d}{j} g(x+j) = c.$$

(Hint: proceed by induction on d.)

5.16 Show that the results in Exercises 5.13 and 5.14 do not hold for the dimension $t = \deg(g) + 1$ since there exists a constant $C_t > 0$ depending only on t such that, with the obvious meaning of $\mathcal{P}_N^{(t)}$, we get

$$D_N(\mathcal{P}_N^{(t)}) \geq C_t \qquad \text{for } 1 \leq N \leq p.$$

(Hint: use Exercise 5.15 as well as Theorem 4.1.41 with a suitable function f.)

5.17 Let p be a prime number and put

$$K(a) = \sum_{c \in \mathbb{F}_p^*} \chi_p(ac^{-1} + c) \qquad \text{for all } a \in \mathbb{F}_p^*.$$

(a) Prove that $K(a)$ is a real number for all $a \in \mathbb{F}_p^*$.

(b) Prove that

$$\sum_{a \in \mathbb{F}_p^*} K(a)^2 = p^2 - p - 1.$$

(c) Deduce from part (b) that the bound in Proposition 5.3.21 is in general best possible up to an absolute constant.

5.18 Let $p \geq 5$ be a prime number and let $(x_n)_{n=0}^{\infty}$ be a sequence of explicit inversive pseudorandom numbers generated by (5.27) and (5.28). Let $\mathcal{P}^{(2)}$ be the point set consisting of the points

$$\mathbf{x}_n = (x_n, x_{n+1}) \in [0,1)^2 \qquad \text{for } n = 0, 1, \ldots, p-1.$$

Prove that the discrepancy $D_p(\mathcal{P}^{(2)})$ of $\mathcal{P}^{(2)}$ satisfies $D_p(\mathcal{P}^{(2)}) = O\left(p^{-1/2} (\log p)^2\right)$ with an absolute implied constant. (Hint: use Proposition 5.3.21.)

5.19 Let $m = 2^k$ with an integer $k \geq 2$ and generate a sequence $(z_n)_{n=0}^{\infty}$ of elements of Z_m by the linear congruential method (5.2).

(a) Show that it is not a good idea to generate a sequence $(b_n)_{n=0}^{\infty}$ of pseudorandom bits by letting b_n be the least significant bit of z_n for all $n \geq 0$.

(b) What can be said if we replace b_n by c_n, where c_n is the coefficient of 2^{k-1} in the binary representation of z_n?

5.20 Prove that if a real number α is normal to the base b for some integer $b \geq 2$, then $m\alpha$ is normal to the base b for every nonzero integer m. (Hint: use the theory of uniformly distributed sequences in Section 4.1.)

5.21 Prove that if a real number α is normal to the base b^k for some integers $b \geq 2$ and $k \geq 2$, then α is normal to the base b.

5.22 Let $f(x) \in \mathbb{F}_q[x]$ be a monic irreducible polynomial over the finite field $\mathbb{F}_q$ of degree d with $f(x) \neq x$. Prove that $f(x)$ is a primitive polynomial over $\mathbb{F}_q$ if and only if $q^d - 1$ is the least positive integer k such that $f(x)$ divides $x^k - 1$ in $\mathbb{F}_q[x]$.

5.23 Let $m(x) = x^4 + x^3 + 1 \in \mathbb{F}_2[x]$.

(a) Verify that $m(x)$ is a primitive polynomial over $\mathbb{F}_2$.

(b) Compute the terms of a corresponding maximal period sequence $\mathcal{B}_4$ over $\mathbb{F}_2$ explicitly.

(c) Verify by a direct computation that $C_h(\mathcal{B}_4) = -1$ for $h = 1, 2, 3, 4$.

5.24 Let $\mathcal{B}_d$ be a maximal period sequence over the finite field $\mathbb{F}_q$ with least period length $q^d - 1$. List the terms in the first period of $\mathcal{B}_d$ in reverse order and continue periodically. Prove that the resulting sequence is again a maximal period sequence over $\mathbb{F}_q$ with least period length $q^d - 1$.

5.25 For every maximal period sequence $(a_n)_{n=0}^{\infty}$ over the finite field $\mathbb{F}_q$ with least period length $q^d - 1$ and for every $r \in \mathbb{N}$ with $\gcd(r, q^d - 1) = 1$, prove that $(a_{rn})_{n=0}^{\infty}$ is again a maximal period sequence over $\mathbb{F}_q$ with least period length $q^d - 1$.

5.26 Let $(a_n)_{n=0}^{\infty}$ be a maximal period sequence over the finite field $\mathbb{F}_q$ with least period length $q^d - 1$. Prove that for every $c \in \mathbb{F}_q^*$, the sequence $(a_n + c)_{n=0}^{\infty}$ has again least period length $q^d - 1$, but it is not a maximal period sequence over $\mathbb{F}_q$.

Chapter 6

Further Applications

The set $\mathbb{Z}$, says a proverb in Finnish,
is infinite and cannot diminish.
Every integer is applicable,
no matter how weird or despicable,
so this book's story will never finish.

6.1 Check-digit systems

6.1.1 Definition and examples

Check-digit systems and error-correcting codes (see Chapter 3 for the latter) are birds of a feather, but it must be conceded that error-correcting codes are the more colorful birds. Just like error-correcting codes, check-digit systems help to eliminate errors in data, but their aims are more modest than those of error-correcting codes. In a check-digit system we extend an identification number, as for example a bank account number, by a control symbol primarily to detect any single error. A check-digit system can be formally defined over any finite abelian group.

Definition 6.1.1 A *check-digit system* over a finite abelian group G (with the additive notation) consists of $n \geq 2$ permutations $f_1, \ldots, f_n$ of G and an element $c \in G$. A word $a_1 \cdots a_{n-1} \in G^{n-1}$ of length $n-1$ is extended to a word of length n by appending to it a check digit a_n such that

$$f_1(a_1) + \cdots + f_n(a_n) = c. \tag{6.1}$$

In practice very often the finite abelian group $G = Z_m$ consisting of the least residue system modulo m with addition modulo m is used (compare with Example 1.3.6).

Example 6.1.2 The *Universal Product Code* (*UPC*) is a barcode widely used in the United States, Canada, and many other countries for tracking trade items in stores. Its most common form UPC-12 consists of 12 decimal digits forming the word $a_1 \cdots a_{12} \in Z_{10}^{12}$, where a_{12} is a check digit chosen such that the control equation (6.1) is given by

$$3(a_1 + a_3 + a_5 + a_7 + a_9 + a_{11}) + a_2 + a_4 + a_6 + a_8 + a_{10} + a_{12} \equiv 0 \pmod{10}.$$

Thus, for the UPC-12 only two types of permutations f_i, $i = 1, \ldots, 12$, of $G = Z_{10}$ are employed in Definition 6.1.1, namely $a \in Z_{10} \mapsto 3a \in Z_{10}$ and the identity map $a \in Z_{10} \mapsto a \in Z_{10}$. Similarly, a *European Article Number* (*EAN*) or *International Article Number* $a_1 \cdots a_{13} \in Z_{10}^{13}$ consists of one more decimal digit and has to satisfy the control equation

$$3(a_2 + a_4 + a_6 + a_8 + a_{10} + a_{12}) + a_1 + a_3 + a_5 + a_7 + a_9 + a_{11} + a_{13} \equiv 0 \pmod{10}.$$

For instance, a package of frozen peas of a well-known Austrian brand bears the EAN

$$9008695928723$$

and it is a nice little exercise in arithmetic modulo 10 to verify the control equation.

Example 6.1.3 The *International Standard Book Number* (*ISBN*) identifies books, as the name suggests. The version ISBN-10 was used until 2007. It starts with one or more leading digits for the language area. For books published in most English-speaking countries, the first digit is either 0 or 1, whereas for German-speaking countries the first digit is 3. The country prefix is followed by digits for publisher and book title and a check digit. An ISBN-10 is given by Definition 6.1.1 with $G = Z_{11}$, $f_i(a) = ia$ for $a \in Z_{11}$ and $i = 1, \ldots, 10$, and furthermore $c = 0$, where the symbol X is used for 10. Since 2007, ISBNs contain 13 decimal digits with an additional prefix 978 or 979. An ISBN-13 $a_1 \cdots a_{13} \in Z_{10}^{13}$ has to satisfy

$$a_1 + a_3 + a_5 + a_7 + a_9 + a_{11} + a_{13} + 3(a_2 + a_4 + a_6 + a_8 + a_{10} + a_{12}) \equiv 0 \pmod{10}.$$

An ISBN-10 can be easily converted to an ISBN-13 by adding the prefix 978 or 979 and calculating the new check digit. For example, the ISBN-10 of the book [53] is 1-4020-5333-9 which is converted to the ISBN-13 978-1-4020-5333-7.

Example 6.1.4 The *International Bank Account Number* (*IBAN*) is used in the European Union and in many countries outside the EU for the purpose of standardizing payments. An IBAN typically comprises 20 to 34 alphanumeric characters. It starts with two letters representing the country code, such as AT for Austria, FI for Finland, and DE for Germany. The country code is followed by two decimal check digits and then by alphanumeric characters specifying the bank and the account number. Checking the validity of an IBAN is more cumbersome than for a UPC, an EAN, or an ISBN. First of all, the first four alphanumeric characters of the IBAN (that is, the country code and the check digits) are moved to the end of the string. Then all letters in the IBAN are converted to integers according to the scheme $\mathrm{A} \leftrightarrow 10, \mathrm{B} \leftrightarrow 11, \ldots, \mathrm{Z} \leftrightarrow 35$. The resulting string of decimal digits is interpreted as the decimal representation of a positive integer N. Finally, it is checked whether this integer N is congruent to 1 modulo 97. This falls into the pattern of the control equation (6.1) with $G = Z_{97}$ and $c = 1$. The permutations f_i of Z_{97} in (6.1) are given by the multiplication modulo 97 of an element of Z_{97} by a suitable power of 10 (in the present case, the f_i are applied only to the elements $0, 1, \ldots, 9$ of Z_{97}). We had planned to offer the IBANs of our secret Swiss bank accounts for practicing the validation of IBANs, but we decided against it at the very last minute. So we are afraid you have to use your own IBAN for this exercise.

6.1.2 Neighbor transpositions and orthomorphisms

The most common errors that should be eliminated by a check-digit system are single errors ($a \mapsto b$ with $b \neq a$), and they are always detected since the f_i in Definition 6.1.1 are permutations. Another type of common errors, occurring particularly in long words such as IBANs, is formed by neighbor transpositions ($ab \mapsto ba$ with $b \neq a$). However, many check-digit systems do not detect this kind of error. For example, the UPC-12 does not detect the neighbor transposition $a_i a_{i+1} \mapsto a_{i+1} a_i$ if $a_i \equiv a_{i+1} \pmod 5$ for some $i = 1, \ldots, 11$.

By the control equation (6.1), the neighbor transposition $a_i a_{i+1} \mapsto a_{i+1} a_i$ (with $i = 1, \ldots, n-1$) is detected if and only if

$$f_i(a_i) + f_{i+1}(a_{i+1}) \neq f_i(a_{i+1}) + f_{i+1}(a_i). \qquad (6.2)$$

Put $a = f_i(a_i)$, $b = f_i(a_{i+1})$, and $F_i = f_{i+1} \circ f_i^{-1}$. Then (6.2) is equivalent to

$$F_i(a) - a \neq F_i(b) - b \qquad \text{for } a, b \in G,\ a \neq b.$$

Therefore F_i must be a permutation of G with the additional property in the following definition.

Definition 6.1.5 A permutation f of the finite abelian group G is an *orthomorphism* of G if $f - \mathrm{id}_G$ is also a permutation of G, where id_G denotes the identity map on G.

Example 6.1.6 The map $f : a \in Z_m \mapsto ca \in Z_m$ with a fixed $c \in Z_m$ is an orthomorphism of Z_m if and only if $\gcd(c, m) = \gcd(c-1, m) = 1$. Such an orthomorphism is called a *linear orthomorphism* of Z_m.

Example 6.1.7 For the ISBN-10 we get $F_i(a) = f_{i+1}(f_i^{-1}(a)) = (i+1)i^{-1}a$ for $i = 1, \ldots, 9$ and all $a \in Z_{11}$. Therefore all F_i are linear orthomorphisms of Z_{11} and all neighbor transpositions are detected.

Theorem 6.1.8 *There exists a check-digit system over the finite abelian group G that detects all neighbor transpositions if and only if there exists an orthomorphism of G.*

Proof. As we have seen, each permutation F_i of G defined above is an orthomorphism of G if all neighbor transpositions are detected. Conversely, let f be an arbitrary orthomorphism of G and define $f_0 = \mathrm{id}_G$ and $f_{i+1} = f \circ f_i$ for $i = 0, 1, \dots, n-1$. Then we get $f_{i+1} \circ f_i^{-1} = f$ and (6.2) is satisfied since f is an orthomorphism of G. Hence all neighbor transpositions are detected. □

In many practical applications of check-digit systems, the finite abelian group G is chosen to be Z_p with a prime number p (see the ISBN-10 in Example 6.1.3 and the IBAN in Example 6.1.4). As we know, we can view Z_p as the finite prime field $\mathbb{F}_p$ (see Theorem 1.4.5 and Remark 1.4.6). Now by Proposition 5.3.2, any self-map of $\mathbb{F}_p$ can be represented by a polynomial over $\mathbb{F}_p$ of degree less than p, and if the considered self-map of $\mathbb{F}_p$ is a permutation of $\mathbb{F}_p$, then the representing polynomial over $\mathbb{F}_p$ is a permutation polynomial of $\mathbb{F}_p$ (see Definition 5.3.4). Hence if G is the additive group $\mathbb{F}_p$, then the maps f_i in Definition 6.1.1 can be taken to be permutation polynomials of $\mathbb{F}_p$. In the examples ISBN-10 and IBAN, these permutation polynomials of $\mathbb{F}_p$ are linear polynomials over $\mathbb{F}_p$. Linear orthomorphisms f of $Z_p = \mathbb{F}_p$ (see Example 6.1.6) are given by $f(x) = cx \in \mathbb{F}_p[x]$ with $c \in \mathbb{F}_p \setminus \{0, 1\}$.

Next we study the slightly more complicated class of *quadratic orthomorphisms* of $\mathbb{F}_p$ in (6.3) below, which are connected with quadratic residues and nonresidues (see Definition 1.2.21).

Proposition 6.1.9 *Let p be an odd prime number and let*

$$f_{a,b}(x) = \frac{a-b}{2} x^{(p+1)/2} + \frac{a+b}{2} x \in \mathbb{F}_p[x] \quad \textit{for } a, b \in \mathbb{F}_p,\ a \neq b. \tag{6.3}$$

If $r \in \mathbb{F}_p$, then

$$f_{a,b}(r) = \begin{cases} ar & \textit{if } r \textit{ is a quadratic residue modulo } p, \\ br & \textit{if } r \textit{ is a quadratic nonresidue modulo } p, \\ 0 & \textit{if } r = 0. \end{cases}$$

Furthermore, $f_{a,b}$ is a permutation polynomial of $\mathbb{F}_p$ if and only if ab is a quadratic residue modulo p and an orthomorphism of $\mathbb{F}_p$ if and only if additionally $(a-1)(b-1)$ is a quadratic residue modulo p.

Proof. The formula for $f_{a,b}(r)$ follows from

$$f_{a,b}(r) = r\Big(\frac{a-b}{2} r^{(p-1)/2} + \frac{a+b}{2}\Big)$$

and Proposition 1.2.23. If ab is a quadratic residue modulo p, then $\left(\frac{a}{p}\right) = \left(\frac{b}{p}\right)$ by Proposition 1.2.24, that is, a and b have the same quadratic-residue behavior modulo p, and then the same proposition and the formula for $f_{a,b}(r)$ show that $f_{a,b}$ is a permutation polynomial of $\mathbb{F}_p$. On the other hand, if $ab = 0 \in \mathbb{F}_p$, then $f_{a,b}$ is clearly not a permutation polynomial of $\mathbb{F}_p$, and if ab is a quadratic nonresidue modulo p, then $f_{a,b}(1) = a = b(b^{-1}a) = f_{a,b}(b^{-1}a)$ since then also $b^{-1}a \neq 1 \in \mathbb{F}_p$ is a quadratic nonresidue modulo p, and so $f_{a,b}$ is not a permutation polynomial of $\mathbb{F}_p$. Finally, we see from (6.3) that

$$f_{a,b}(x) - x = \frac{a-b}{2}x^{(p+1)/2} + \Big(\frac{a+b}{2} - 1\Big)x = f_{a-1,b-1}(x),$$

and so $f_{a,b}$ is an orthomorphism of $\mathbb{F}_p$ if and only if both ab and $(a-1)(b-1)$ are quadratic residues modulo p. □

Theorem 6.1.10 *Let p be an odd prime number. Then the number of ordered pairs $(a,b) \in \mathbb{F}_p^2$ with $a \neq b$ such that $f_{a,b}$ is a permutation polynomial of $\mathbb{F}_p$ is*

$$\frac{(p-1)(p-3)}{2}.$$

The number of ordered pairs $(a,b) \in \mathbb{F}_p^2$ with $a \neq b$ such that $f_{a,b}$ is an orthomorphism of $\mathbb{F}_p$ is

$$\frac{(p-3)(p-5)}{4}.$$

Proof. We take any of the $p-1$ possibilities for $a \in \mathbb{F}_p^*$ and choose $b \in \mathbb{F}_p$ with $b \neq a$ such that ab is a quadratic residue modulo p (see Proposition 6.1.9). According to Remark 1.2.26, there are exactly $(p-1)/2 - 1 = (p-3)/2$ choices for b.

For counting the orthomorphisms $f_{a,b}$ of $\mathbb{F}_p$, we can assume by Proposition 6.1.9 that $a, b \in \mathbb{F}_p^*$. The number of orthomorphisms $f_{a,b}$ of $\mathbb{F}_p$ with $a \neq b$ equals the number of orthomorphisms $f_{a,ab}$ of $\mathbb{F}_p$ with $b \neq 1$. Now $f_{a,ab}$ is an orthomorphism of $\mathbb{F}_p$ if and only if a^2b and $(a-1)(ab-1) = b(a-1)(a-b^{-1})$ are both quadratic residues modulo p, which is true if and only if b and $(a-1)(a-b^{-1})^{-1}$ are both quadratic residues modulo p. The number of quadratic residues $b = 2, \ldots, p-1$ modulo p is $(p-3)/2$. If a runs through all elements of $\mathbb{F}_p^* \setminus \{1, b^{-1}\}$, then $(a-1)(a-b^{-1})^{-1}$ runs through all elements of $\mathbb{F}_p^* \setminus \{1, b\}$. So for any fixed b it runs through $(p-1)/2 - 2 = (p-5)/2$ different quadratic residues modulo p, and the result follows. □

Remark 6.1.11 We deduce from Theorem 6.1.10 that for a large prime number p, a random choice of $(a,b) \in \mathbb{F}_p^2$ yields a permutation polynomial $f_{a,b}$ of $\mathbb{F}_p$, respectively an orthomorphism $f_{a,b}$ of $\mathbb{F}_p$, with probability about $\frac{1}{2}$, respectively about $\frac{1}{4}$.

6.1.3 Permutations for detecting other frequent errors

The detection of several other types of frequent errors is guaranteed only if the permutations $f_1, \dots, f_n$ of G in Definition 6.1.1 satisfy additional conditions. Besides single errors and neighbor transpositions, these types of errors are:

- jump transpositions $acb \mapsto bca$ (with $b \neq a$);
- twin errors $aa \mapsto bb$ (with $b \neq a$);
- jump twin errors $aca \mapsto bcb$ (with $b \neq a$).

We examine the case where $f_i = f^{(i)}$ for $i = 1, \dots, n$ is the ith iterate of a fixed permutation f of G, that is, $f_0 = \mathrm{id}_G$ and $f_{i+1} = f \circ f_i$ for $i = 0, 1, \dots, n-1$. All single errors are detected since f is a permutation of G, and by the proof of Theorem 6.1.8 all neighbor transpositions are detected if and only if $f - \mathrm{id}_G$ is a permutation of G. It is easy to see that all twin errors, jump transpositions, and jump twin errors are detected whenever $f + \mathrm{id}_G$, $f_2 - \mathrm{id}_G$, and $f_2 + \mathrm{id}_G$, respectively, is a permutation of G.

Example 6.1.12 Let p be an odd prime number. Consider the self-map f of $\mathbb{F}_p$ represented by the polynomial $f(x) = cx \in \mathbb{F}_p[x]$ with $c \in \mathbb{F}_p$. Then all three polynomials $f(x)$ and $f(x) \pm x$ are permutation polynomials of $\mathbb{F}_p$ if and only if $c \notin \{0, 1, -1\}$. With the notation above, we get $f_2(x) = f(f(x)) = c^2x$, and so $f_2(x) - x = (c^2 - 1)x$ is a permutation polynomial of $\mathbb{F}_p$ if and only if $c \notin \{1, -1\}$. Furthermore, $f_2(x) + x = (c^2 + 1)x$ is a permutation polynomial of $\mathbb{F}_p$ if and only if $c^2 \neq -1$. We recall from Example 1.2.25 that -1 is a quadratic residue modulo p if and only if $p \equiv 1 \pmod 4$. Now we inspect for which values of $c \in \mathbb{F}_p$ all five polynomials $f(x)$, $f(x) \pm x$, and $f_2(x) \pm x$ are permutation polynomials of $\mathbb{F}_p$. If $p \equiv 3 \pmod 4$, then this happens precisely for all $c \notin \{0, 1, -1\}$, and so there are exactly $p - 3$ choices for c. If $p \equiv 1 \pmod 4$, then we have to exclude also the two roots $d \in \mathbb{F}_p$ and $-d \in \mathbb{F}_p$ of $x^2 + 1 \in \mathbb{F}_p[x]$, and so there are exactly $p - 5$ choices for c.

Example 6.1.13 We see from Example 6.1.3 that ISBN-10 does not detect errors of the form $a_5a_6 \mapsto (a_5+b)(a_6+b)$ with $b \in Z_{11} \setminus \{0\}$, including twin errors, at positions 5 and 6 for instance. After the fixed coordinate permutation $a_i \mapsto a_{2^i \pmod{11}}$ for $i = 1, \dots, 10$, the modified ISBN-10 can be considered a check-digit system over $\mathbb{F}_{11}$ defined by $f(x) = 2x \in \mathbb{F}_{11}[x]$ and $f_i = f^{(i)}$ for $i = 1, \dots, 10$. Hence by Example 6.1.12, the modified version of ISBN-10 detects all five types of errors mentioned above.

We recall from the discussion above that a check-digit system based on the iterates $f_i = f^{(i)}$ of a fixed permutation f of G corrects all twin errors whenever

$f + \mathrm{id}_G$ is also a permutation of G. This property is captured by the following definition.

Definition 6.1.14 A permutation f of the finite abelian group G is a *complete mapping* of G if $f + \mathrm{id}_G$ is also a permutation of G.

Remark 6.1.15 A permutation f of G is a complete mapping of G if and only if $-f$ is an orthomorphism of G. A check-digit system defined by the permutation f of G and $f_i = f^{(i)}$ for $i = 1, \ldots, n$ can detect all single errors, neighbor transpositions, twin errors, jump transpositions, and jump twin errors whenever f and f_2 are both complete mappings and orthomorphisms of G. According to a definition given in Evans [48], f is a *strong complete mapping* of G if f is both a complete mapping and an orthomorphism of G.

Remark 6.1.16 Complete mappings are also pertinent to the construction of orthogonal latin squares. We refer to Remark 4.4.13 for the definition of latin squares and orthogonal latin squares of order $b \geq 2$. Let f be a complete mapping of Z_b. Then we claim that $S_1 = (a_{ij})_{1 \leq i,j \leq b}$ with $a_{ij} \equiv i + j \pmod{b}$ and $S_2 = (c_{ij})_{1 \leq i,j \leq b}$ with $c_{ij} \equiv f(j) - i \pmod{b}$ are orthogonal latin squares of order b. It is trivial that S_1 is a latin square, and S_2 is a latin square since f is a permutation of Z_b. Now assume that $(a_{ij}, c_{ij}) = (a_{k\ell}, c_{k\ell})$, or equivalently

$$i + j \equiv k + \ell \pmod{b} \quad \text{and} \quad f(j) - i \equiv f(\ell) - k \pmod{b}.$$

Adding these congruences yields $f(j) + j \equiv f(\ell) + \ell \pmod{b}$ and thus $j = \ell$ since f is a complete mapping of Z_b. Then also $i = k$ and the result follows. Orthogonal latin squares have many applications, for instance to the design of agricultural experiments (see [94, Section 1.4 and Chapter 16]). The authors once attended a talk on this topic with the funny title “Applications of finite fields to fields”.

We consider again the polynomials $f_{a,b} \in \mathbb{F}_p[x]$ defined by (6.3). In Theorem 6.1.20 below we prove an asymptotic formula for the number of $(a, b) \in \mathbb{F}_p^2$ with $a \neq b$ such that the five polynomials $f(x)$, $f(x) \pm x$, and $f(f(x)) \pm x$ with $f = f_{a,b}$ are all permutation polynomials of $\mathbb{F}_p$ and can thus be used to design check-digit systems that detect all the above five types of frequent errors. Let $\left(\frac{a}{p}\right)$ be the Legendre symbol introduced in Definition 1.2.22.

Corollary 6.1.17 *Let p be an odd prime number, let $a, b \in \mathbb{F}_p$ with $a \neq b$, and let $f_{a,b}(x) \in \mathbb{F}_p[x]$ be defined by (6.3). Then the three polynomials $f_{a,b}(x)$ and $f_{a,b}(x) \pm x$ are all permutation polynomials of $\mathbb{F}_p$ if and only if $a, b \notin \{-1, 0, 1\}$,*

$$\left(\frac{a}{p}\right) = \left(\frac{b}{p}\right), \quad \left(\frac{a-1}{p}\right) = \left(\frac{b-1}{p}\right), \quad \textit{and} \left(\frac{a+1}{p}\right) = \left(\frac{b+1}{p}\right). \qquad (6.4)$$

Proof. This follows immediately from Proposition 6.1.9. □

Lemma 6.1.18 *Let p be an odd prime number, let $a, b \in \mathbb{F}_p$ with $a \neq b$, and let $f_{a,b}(x) \in \mathbb{F}_p[x]$ be defined by (6.3). If $\left(\frac{a}{p}\right) = \left(\frac{b}{p}\right) \neq 0$, then for $r \in \mathbb{F}_p^*$,*

$$f_{a,b}(f_{a,b}(r)) = \begin{cases} a^2 r & \textit{if } \left(\frac{a}{p}\right) = \left(\frac{r}{p}\right) = 1, \\ b^2 r & \textit{if } \left(\frac{a}{p}\right) = -\left(\frac{r}{p}\right) = 1, \\ abr & \textit{if } \left(\frac{a}{p}\right) = -1. \end{cases}$$

Proof. We distinguish four cases, according to the four possible combinations of values of $\left(\frac{a}{p}\right)$ and $\left(\frac{r}{p}\right)$. If $\left(\frac{a}{p}\right) = \left(\frac{r}{p}\right) = 1$, then by Proposition 6.1.9 we get $f_{a,b}(r) = ar$. Furthermore $\left(\frac{ar}{p}\right) = 1$, and so another application of Proposition 6.1.9 yields $f_{a,b}(f_{a,b}(r)) = f_{a,b}(ar) = a^2 r$. The other three cases are treated in an analogous way. □

Lemma 6.1.19 *Let p be an odd prime number, let $a, b \in \mathbb{F}_p^*$ with*

$$a \neq b \quad \textit{and} \quad a^2, b^2 \notin \{-1, 1\}, \tag{6.5}$$

and let $f_{a,b}(x) \in \mathbb{F}_p[x]$ be defined by (6.3). If the three polynomials $f_{a,b}(x)$ and $f_{a,b}(x) \pm x$ are permutation polynomials of $\mathbb{F}_p$, then the two polynomials $f_{a,b}(f_{a,b}(x)) \pm x$ are permutation polynomials of $\mathbb{F}_p$ if and only if

$$\left(\frac{a^2+1}{p}\right) = \left(\frac{b^2+1}{p}\right) \quad \textit{and} \quad \left(\frac{a}{p}\right) = 1 \tag{6.6}$$

or

$$\left(\frac{a}{p}\right) = -1. \tag{6.7}$$

Proof. If $\left(\frac{a}{p}\right) = 1$, then Lemma 6.1.18 shows that for $r \in \mathbb{F}_p$,

$$f_{a,b}(f_{a,b}(r)) = \begin{cases} a^2 r & \text{if } \left(\frac{r}{p}\right) = 1, \\ b^2 r & \text{otherwise.} \end{cases}$$

Hence $f_{a,b}(f_{a,b}(x)) \pm x$ are both permutation polynomials of $\mathbb{F}_p$ if and only if

$$\left(\frac{a^2-1}{p}\right) = \left(\frac{b^2-1}{p}\right) \quad \text{and} \quad \left(\frac{a^2+1}{p}\right) = \left(\frac{b^2+1}{p}\right).$$

The first condition is already covered by (6.4). If $\left(\frac{a}{p}\right) = -1$, then the result follows immediately from Lemma 6.1.18. □

Theorem 6.1.20 *Let p be an odd prime number and let $f_{a,b}(x) \in \mathbb{F}_p[x]$ be defined by (6.3). Let N be the number of ordered pairs $(a,b) \in \mathbb{F}_p^2$ with $a \neq b$ such that the five polynomials $f_{a,b}(x)$, $f_{a,b}(x) \pm x$, and $f_{a,b}(f_{a,b}(x)) \pm x$ are all permutation polynomials of $\mathbb{F}_p$. Then*

$$N = \frac{3p^2}{32} + O(p),$$

where the implied constant is absolute.

Proof. Let N_1 and N_2 be the numbers of ordered pairs (a,b) with $a, b \in \mathbb{F}_p^*$ satisfying

$$(6.4),\ (6.5),\quad \text{and}\quad (6.6),$$

$$(6.4),\ (6.5),\quad \text{and}\quad (6.7),$$

respectively. Then $N = N_1 + N_2$.

For typographic convenience we now write $\eta(a) = \left(\frac{a}{p}\right)$ for $a \in \mathbb{F}_p$, that is, η is the quadratic character of $\mathbb{F}_p$ (see Remark 1.4.53) with the additional stipulation $\eta(0) = 0$. Then for $a, b \in \mathbb{F}_p^*$ satisfying (6.5) we obtain

$$\begin{aligned}\Delta_1(a,b) &:= \frac{1}{32}(1+\eta(a))(1+\eta(b))(1+\eta((a-1)(b-1)))\\ &\qquad (1+\eta((a+1)(b+1)))(1+\eta((a^2+1)(b^2+1)))\\ &= \begin{cases} 1 & \text{if } (a,b) \text{ satisfies (6.4) and (6.6)},\\ 0 & \text{otherwise},\end{cases}\end{aligned}$$

and so

$$N_1 = \sum_{a,b\in\mathbb{F}_p} \Delta_1(a,b) + O(p),$$

since the number of ordered pairs $(a,b) \in \mathbb{F}_p^2$ that do not satisfy (6.5) or with $ab = 0$ is $O(p)$. Hence we get

$$N_1 = \frac{1}{32} \sum_{j_1,j_2,j_3,j_4,j_5=0}^{1} S_{j_1,j_2,j_3,j_4,j_5} + O(p),$$

where

$$\begin{aligned}S_{j_1,j_2,j_3,j_4,j_5} := &\sum_{a\in\mathbb{F}_p} \eta(a^{j_1}(a-1)^{j_3}(a+1)^{j_4}(a^2+1)^{j_5})\\ &\qquad \sum_{b\in\mathbb{F}_p} \eta(b^{j_2}(b-1)^{j_3}(b+1)^{j_4}(b^2+1)^{j_5})\end{aligned}$$

with the convention $0^0 = 1 \in \mathbb{F}_p$. We note that $S_{0,0,0,0,0} = p^2$, and furthermore

$$S_{1,0,0,0,0} = p \sum_{a\in\mathbb{F}_p} \eta(a) = 0 \quad \text{and} \quad S_{0,1,0,0,0} = p \sum_{b\in\mathbb{F}_p} \eta(b) = 0$$

by Example 1.3.35.

In Proposition 5.3.8 we formulated the Weil bound for additive characters of $\mathbb{F}_p$. There is also a Weil bound for multiplicative characters of $\mathbb{F}_p$ (see [102, Theorem 5.41]), and we use the following special case thereof: if $f(x) \in \mathbb{F}_p[x]$ is a monic polynomial of positive degree which is not a square of another polynomial, then

$$\Big|\sum_{c\in\mathbb{F}_p} \eta(f(c))\Big| \leq (\deg(f)-1)p^{1/2}. \tag{6.8}$$

In the remaining $O(1)$ cases, both monic polynomials

$$x^{j_1}(x-1)^{j_3}(x+1)^{j_4}(x^2+1)^{j_5} \text{ and } x^{j_2}(x-1)^{j_3}(x+1)^{j_4}(x^2+1)^{j_5}$$

are not squares and we can apply the Weil bound to the sums over a and b to get

$$S_{j_1,j_2,j_3,j_4,j_5} = O(p).$$

Collecting everything we obtain

$$N_1 = \frac{p^2}{32} + O(p)$$

with an absolute implied constant.

Next we observe that

$$N_2 = \sum_{a,b\in\mathbb{F}_p} \Delta_2(a,b) + O(p),$$

where

$$\Delta_2(a,b) = \frac{1}{16}(1-\eta(a))(1-\eta(b))(1+\eta((a-1)(b-1)))(1+\eta((a+1)(b+1))).$$

It follows that

$$N_2 = \frac{1}{16}\sum_{j_1,j_2,j_3,j_4=0}^{1} S_{j_1,j_2,j_3,j_4} + O(p),$$

where

$$S_{j_1,j_2,j_3,j_4} := (-1)^{j_1+j_2}\sum_{a\in\mathbb{F}_p}\eta(a^{j_1}(a-1)^{j_3}(a+1)^{j_4})\sum_{b\in\mathbb{F}_p}\eta(b^{j_2}(b-1)^{j_3}(b+1)^{j_4}).$$

We note that $S_{0,0,0,0} = p^2$ and $S_{1,0,0,0} = S_{0,1,0,0} = 0$. In the remaining cases, we can apply the Weil bound (6.8) and we get

$$N_2 = \frac{p^2}{16} + O(p)$$

with an absolute implied constant, which finishes the proof. □

Remark 6.1.21 If p is large, then Theorem 6.1.20 shows that the probability that $f_{a,b}(x)$, $f_{a,b}(x) \pm x$, and $f_{a,b}(f_{a,b}(x)) \pm x$ are all permutation polynomials of $\mathbb{F}_p$ for randomly chosen $a, b \in \mathbb{F}_p$ with $a \neq b$ is close to $\frac{3}{32}$.

6.2 Covering sets and packing sets

6.2.1 Covering sets and rewriting schemes

A flash memory is an electronic storage medium that can be erased and rewritten. Erasures can be performed only on a blockwise basis, and this limitation of flash memories leads to interesting number-theoretic problems.

We consider a set of n flash memory cells, each capable of storing an element of the finite prime field $\mathbb{F}_p$, and a set $\mathcal{S} = \{s_1, \dots, s_n\} \subseteq \mathbb{F}_p$ of size n. We identify $\mathcal{S}$ with the vector $\mathbf{s} = (s_1, \dots, s_n) \in \mathbb{F}_p^n$. We store a value $v \in \mathbb{F}_p$ in the n memory cells by first choosing a vector $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{F}_p^n$ for which the dot product $\mathbf{x} \cdot \mathbf{s} = x_1 s_1 + \cdots + x_n s_n$ satisfies $\mathbf{x} \cdot \mathbf{s} = v$ and then storing x_i in the ith cell for $1 \leq i \leq n$. For rewriting v by some $v' \in \mathbb{F}_p$, we have to choose $\mathbf{x}' = (x_1', \dots, x_n') \in \mathbb{F}_p^n$ with $\mathbf{x}' \cdot \mathbf{s} = v'$ and $x_i' \in \{x_i - \mu, x_i - \mu + 1, \dots, x_i + \lambda\}$ due to the limitation of flash memory mentioned above, where λ and μ are prescribed small nonnegative integers. For the sake of efficiency, we want to leave as many cells as possible unchanged, and in the extreme case we allow only a single cell to change. These considerations lead to the following concept.

Definition 6.2.1 Let p be a prime number. For a given set $\mathcal{M} = \{-\mu, -\mu + 1, \dots, \lambda\} \setminus \{0\}$ with $\lambda, \mu \in Z_p = \mathbb{F}_p$ not both 0, a nonempty subset $\mathcal{S}$ of $Z_p = \mathbb{F}_p$ is called a $(\lambda, \mu; p)$-*covering set* if

$$\mathcal{M}\mathcal{S} := \{ms \in \mathbb{F}_p : m \in \mathcal{M},\ s \in \mathcal{S}\} = \mathbb{F}_p.$$

If $\mathcal{S}$ is a $(\lambda, \mu; p)$-covering set, then with the notation above we can write $v' - v = ms_i$ for some $m \in \mathcal{M}$ and $s_i \in \mathcal{S}$, and so it suffices to change x_i to $x_i + m$ to derive $\mathbf{x}'$ from $\mathbf{x}$. Again for the sake of efficiency, we are interested in $(\lambda, \mu; p)$-covering sets of smallest possible size. The lower bound

$$|\mathcal{S}| \geq \left\lceil \frac{p}{\lambda + \mu} \right\rceil \tag{6.9}$$

holds for every $(\lambda, \mu; p)$-covering set $\mathcal{S}$, since obviously $p = |\mathcal{M}\mathcal{S}| \leq (\lambda + \mu)|\mathcal{S}|$. For very small values of λ and μ, it is easy to find covering sets with equality in (6.9).

Example 6.2.2 The case $\lambda + \mu = 1$ is trivial since then $\mathcal{S} = Z_p$ is a $(\lambda, \mu; p)$-covering set meeting the bound (6.9). Let p be a prime number with $p \equiv \pm 3 \pmod 8$. Then 2 is a quadratic nonresidue modulo p according to [152, Theorem 3.3]. It is easily checked that the set of quadratic residues modulo p in Z_p together with $0 \in Z_p$ forms a $(2, 0; p)$-covering set of minimal size $(p + 1)/2$. If $p \equiv 3 \pmod 4$, then -1 is a quadratic nonresidue modulo p by Example 1.2.25. Then the set of quadratic residues modulo p in Z_p together with $0 \in Z_p$ yields a $(1, 1; p)$-covering set of minimal size $(p + 1)/2$.

Although the approach in Example 6.2.2 can be extended to higher-order residues such as cubic residues modulo p (compare with Exercise 1.28), it provides small covering sets only for very small values of λ and μ. However, there is a general construction due to Chen, Shparlinski, and Winterhof [22] which is best possible up to a multiplicative constant.

Theorem 6.2.3 *For all prime numbers $p \geq 3$ and all $\lambda, \mu \in Z_p$ with $max(\lambda, \mu) \geq 2$, there is a $(\lambda, \mu; p)$-covering set $\mathcal{S}$ with*

$$|\mathcal{S}| = 2\left\lceil (p-1)/\max(\lambda, \mu)\right\rceil + 1.$$

Proof. Note that whenever $\mathcal{S}$ is a $(\lambda, \mu; p)$-covering set, then so is $-\mathcal{S} = \{-s : s \in \mathcal{S}\}$. Hence we may restrict ourselves to the case $\lambda \geq \mu$ and we note that $\{1, \ldots, \lambda\} \subseteq \mathcal{M}$.

Put $H = \lceil (p-1)/\lambda \rceil$ and

$$\mathcal{S} = \{\pm j^{-1} \in \mathbb{F}_p : 1 \leq j \leq H\} \cup \{0\}.$$

Note that $\pm j^{-1} = \pm k^{-1}$ for $1 \leq j, k \leq H$ precisely if $k = \pm j$ in $\mathbb{F}_p$. Since $H \leq (p-1)/2$, this can hold only if the plus sign applies and $k = j$. Therefore

$$|\mathcal{S}| = 2H + 1 = 2\lceil (p-1)/\lambda \rceil + 1.$$

Let $a \in \mathbb{F}_p$ be arbitrary. We want to show that $a = ms$ in $\mathbb{F}_p$ for some $m \in \{1, \ldots, \lambda\}$ and $s \in \mathcal{S}$. For $a = 0$ we take $m = 1$ and $s = 0$. If $a \neq 0$, then we consider a as an integer in $\{1, \ldots, p-1\}$. We form the $H+1$ distinct least residues modulo p of the integers ca with $c = 0, 1, \ldots, H$. We partition the interval $[0, p-1]$ into the H disjoint intervals

$$\Big[0, \frac{p-1}{H}\Big], \Big(\frac{p-1}{H}, \frac{2(p-1)}{H}\Big], \ldots, \Big(\frac{(H-1)(p-1)}{H}, p-1\Big].$$

By the pigeon-hole principle, one of these intervals must contain at least two of the considered least residues modulo p. Hence there exist $c_1, c_2 \in \{0, 1, \ldots, H\}$ with $c_1 \neq c_2$ such that the least residue r_1 of $c_1 a$ modulo p and the least residue r_2 of $c_2 a$ modulo p satisfy

$$1 \leq r_1 - r_2 \leq \frac{p-1}{H} \leq \lambda.$$

Now

$$r_1 - r_2 \equiv c_1 a - c_2 a \equiv (c_1 - c_2)a \pmod{p},$$

and so

$$a = (r_1 - r_2)(c_1 - c_2)^{-1} \in \mathbb{F}_p. \tag{6.10}$$

Since $r_1 - r_2 \in \{1, \ldots, \lambda\}$ and $(c_1 - c_2)^{-1} \in \mathcal{S}$, we are done. □

Remark 6.2.4 In the symmetric case $\lambda = \mu$ of Theorem 6.2.3, we can improve the result to $|\mathcal{S}| = \lceil (p-1)/\lambda \rceil + 1$. We proceed as in the proof of Theorem 6.2.3, but we put

$$\mathcal{S} = \{j^{-1} \in \mathbb{F}_p : 1 \leq j \leq H\} \cup \{0\}.$$

After the application of the pigeon-hole principle we can choose $c_1 > c_2$ and we write $1 \leq |r_1 - r_2| \leq \lambda$. Then again (6.10) holds, now with $r_1 - r_2 \in \mathcal{M} = \{-\lambda, -\lambda+1, \dots, \lambda\} \setminus \{0\}$ and $(c_1 - c_2)^{-1} \in \mathcal{S}$.

6.2.2 Packing sets and limited-magnitude error correction

Here is a related concept which is of relevance for communication channels in which only errors of limited magnitude occur.

Definition 6.2.5 Let p be a prime number. For a given set $\mathcal{M} = \{-\mu, -\mu + 1, \dots, \lambda\} \setminus \{0\}$ with $\lambda, \mu \in Z_p = \mathbb{F}_p$ not both 0, a nonempty subset $\mathcal{S}$ of $Z_p = \mathbb{F}_p$ is a $(\lambda, \mu; p)$-*packing set* if

$$|\mathcal{M}\mathcal{S}| = |\mathcal{M}||\mathcal{S}|.$$

Remark 6.2.6 In a $(\lambda, \mu; p)$-limited-magnitude error channel, an element $a \in \mathbb{F}_p$ may be changed into any element $a+e \in \mathbb{F}_p$ with $e \in \mathcal{M} = \{-\mu, -\mu+1, \dots, \lambda\} \setminus \{0\}$. For a set $\mathcal{S} = \{s_1, \dots, s_n\} \subseteq \mathbb{F}_p$ with $n \geq 2$, we define the linear code

$$C = \{(c_1, \dots, c_n) \in \mathbb{F}_p^n : c_1 s_1 + \cdots + c_n s_n = 0\}.$$

If a single error $e \in \mathcal{M}$ occurs at position j, that is, we receive $(v_1, \dots, v_n) = (c_1, \dots, c_j + e, \dots, c_n)$, then we get the syndrome (see Definition 3.2.49)

$$\sum_{i=1}^{n} v_i s_i = e s_j.$$

Hence the set of possible syndromes is $\mathcal{M}\mathcal{S}$. If $\mathcal{S}$ is a $(\lambda, \mu; p)$-packing set, then the syndromes are distinct and C can correct any single limited-magnitude error $e \in \mathcal{M}$ since the syndrome uniquely determines e and j.

Since any nonempty subset of a $(\lambda, \mu; p)$-packing set is again a $(\lambda, \mu; p)$-packing set, we are mainly interested in large packing sets. For every $(\lambda, \mu; p)$-packing set $\mathcal{S}$, the inequality $p \geq |\mathcal{M}\mathcal{S}| = (\lambda + \mu)|\mathcal{S}|$ holds, and thus

$$|\mathcal{S}| \leq \left\lfloor \frac{p}{\lambda + \mu} \right\rfloor.$$

Example 6.2.7 In analogy with Example 6.2.2, we get the following packing sets. If p is a prime number with $p \equiv \pm 3 \pmod 8$, then the set of quadratic residues modulo p in Z_p is a $(2, 0; p)$-packing set of maximal size $(p-1)/2$. If $p \equiv 3 \pmod 4$, then the set of quadratic residues modulo p in Z_p is a $(1, 1; p)$-packing set of maximal size $(p-1)/2$.

Again this approach works only for very small values of λ and μ. A general, but not optimal construction is given in the next result.

Proposition 6.2.8 *Let p be a prime number and let $\lambda, \mu \in Z_p = \mathbb{F}_p$ with $1 \leq \lambda + \mu < p$. Then*

$$\mathcal{S} = \left\{ 1 + j(\lambda + \mu + 1) \in Z_p = \mathbb{F}_p : j = 0, 1, \ldots, \left\lfloor \frac{p - \lambda - \mu - 1}{(\lambda + \mu)(\lambda + \mu + 1)} \right\rfloor \right\}$$

is a $(\lambda, \mu; p)$-packing set.

Proof. Let $\mathcal{M}$ be as in Definition 6.2.5 and assume that $m_1 s_1 = m_2 s_2$ in $\mathbb{F}_p$, that is, $m_1 s_1 \equiv m_2 s_2 \pmod{p}$, for some $m_1, m_2 \in \mathcal{M}$ and $s_1, s_2 \in \mathcal{S}$. Since $s < p/(\lambda + \mu)$ for all $s \in \mathcal{S}$, we get $-\mu \frac{p}{\lambda+\mu} < ms < \lambda \frac{p}{\lambda+\mu}$ for all $m \in \mathcal{M}$ and $s \in \mathcal{S}$, and so $m_1 s_1 = m_2 s_2 \in \mathbb{Z}$. Therefore

$$m_1 \equiv m_1 s_1 \equiv m_2 s_2 \equiv m_2 \pmod{\lambda + \mu + 1},$$

which implies $m_1 = m_2$ and thus $s_1 = s_2$. Hence $\mathcal{S}$ is a $(\lambda, \mu; p)$-packing set. □

6.3 Waring's problem for finite fields

6.3.1 Waring's problem

In 1770 Edward Waring conjectured the following in his *Meditationes Algebraicae*: each positive integer is the sum of at most nine cubes, 19 fourth powers, and so on. It was earlier conjectured by Bachet in the 17th century that each positive integer is the sum of at most four squares. This led to the following definition.

Definition 6.3.1 For every integer $k \geq 2$, let $g(k)$ be the smallest number s of summands such that for each integer $n \geq 1$ there exist integers $h_1, \ldots, h_s \geq 0$ with

$$h_1^k + \cdots + h_s^k = n.$$

The problem of determining $g(k)$ is called *Waring's problem* (for integers). Actually, it is not evident that $g(k)$ is always finite, but this was proved by the mathematical all-rounder David Hilbert (1862–1943) who is famous in particular for the list of 23 problems that he presented at the International Congress of Mathematicians in Paris in 1900. This result of Hilbert that $g(k) < \infty$ for all $k \geq 2$ was classified as one of the three pearls of number theory by Khinchin [77]. We easily get a lower bound on $g(k)$.

Proposition 6.3.2 *The bound*

$$g(k) \geq 2^k + \lfloor (3/2)^k \rfloor - 2 \tag{6.11}$$

holds for all integers $k \geq 2$.

Proof. The integer

$$n = (\lfloor (3/2)^k \rfloor - 1)2^k + (2^k - 1)1^k$$

is smaller than 3^k and therefore has to be represented as a sum of summands 1^k and 2^k only. The $2^k - 1$ summands 1^k cannot be substituted by a summand 2^k, and so the representation is minimal and takes $2^k + \lfloor (3/2)^k \rfloor - 2$ summands. □

It is conjectured that we always have equality in (6.11). This was proved for all sufficiently large k by Mahler [109] and it was verified for a large finite range of values of k in [90]. In particular, it is known that $g(2) = 4$ (a celebrated result of Lagrange, the four-square theorem), $g(3) = 9$, and $g(4) = 19$, as predicted by Bachet and Waring. A detailed discussion of Waring's problem for integers can be found in the survey article [197].

Analogs of Waring's problem can be stated for any ring. In particular, Waring's problem for finite fields studied below has several applications.

Definition 6.3.3 For a positive integer k and a prime power q, the *Waring number* $g(k, q)$ over the finite field $\mathbb{F}_q$ is the smallest number s of summands such that every element $b \in \mathbb{F}_q$ is a sum of s kth powers in $\mathbb{F}_q$, that is, there exist $a_1, \ldots, a_s \in \mathbb{F}_q$ with

$$a_1^k + \cdots + a_s^k = b.$$

If there is an element $b \in \mathbb{F}_q$ that cannot be represented as a sum of kth powers in $\mathbb{F}_q$, then we put $g(k, q) = \infty$.

Lemma 6.3.4 *Let $k \in \mathbb{N}$, let q be a prime power, and put $d = \gcd(k, q-1)$. Then*

$$g(k, q) = g(d, q).$$

Proof. It suffices to show that $\{a^k : a \in \mathbb{F}_q^*\} = \{c^d : c \in \mathbb{F}_q^*\}$. First we can write $a^k = (a^{k/d})^d$ for $a \in \mathbb{F}_q^*$, and so the first set is contained in the second set. Now by Proposition 1.1.5, there exist integers u and v with $d = ku + (q-1)v$, and thus $c^d = (c^u)^k (c^{q-1})^v = (c^u)^k$ for $c \in \mathbb{F}_q^*$ by Proposition 1.4.13, which completes the proof. □

In view of Lemma 6.3.4, we can now restrict the discussion to the case where k divides $q - 1$. The following example shows in particular that we can have $g(k, q) = \infty$, as opposed to Waring's problem for integers where $g(k) < \infty$ for all $k \geq 2$ thanks to Hilbert.

Example 6.3.5 Let us start with the trivial positive divisors k of $q-1$. For $k=1$ it is obvious that $g(1,q)=1$. Now let $k=q-1$. If q is a prime number p, then

$$g(p-1,p)=p-1$$

since $\{a^{p-1} : a \in \mathbb{F}_p\} = \{0,1\}$ and $b = \underbrace{1+\cdots+1}_{b \text{ summands}}$ for $1 \le b \le p-1$. If $q=p^r$ with an integer $r \ge 2$, then $\{a^{q-1} : a \in \mathbb{F}_q\} = \{0,1\} \subseteq \mathbb{F}_p$, and so the elements of $\mathbb{F}_q \setminus \mathbb{F}_p$ cannot be represented as sums of $(q-1)$st powers of elements of $\mathbb{F}_q$. This means by Definition 6.3.3 that $g(q-1,q)=\infty$. Now we consider $k=(q-1)/2$, where q is a power of an odd prime. If q is a prime number $p \ge 3$, then

$$g\Big(\frac{p-1}{2},p\Big) = \frac{p-1}{2}$$

since $\{a^{(p-1)/2} : a \in \mathbb{F}_p\} = \{-1,0,1\}$ and $b = \underbrace{1+\cdots+1}_{b \text{ summands}} = \underbrace{-1-\cdots-1}_{p-b \text{ summands}}$ in $\mathbb{F}_p$ for $1 \le b \le p-1$. If $q=p^r$ with an integer $r \ge 2$, then $\{a^{(q-1)/2} : a \in \mathbb{F}_q\} = \{-1,0,1\} \subseteq \mathbb{F}_p$, and so $g((q-1)/2,q)=\infty$.

The following theorem characterizes the cases where $g(k,q) < \infty$. Actually, it is more transparent to formulate the characterization for the opposite case $g(k,q)=\infty$.

Theorem 6.3.6 *Let $q=p^r$ with a prime number p and an integer $r \ge 1$ and let k be a positive divisor of $q-1$. Then $g(k,q)=\infty$ if and only if $(q-1)/(p^d-1)$ divides k for some proper divisor d of r.*

Proof. It is obvious that the subset

$$B_k := \{a_1^k + \cdots + a_s^k : a_1,\ldots,a_s \in \mathbb{F}_q,\ s=1,2,\ldots\}$$

of $\mathbb{F}_q$ is closed under addition and multiplication. For every fixed $b \in B_k$ with $b \ne 0$, the elements bc with $c \in B_k$ run again through B_k, and since $1 \in B_k$, it follows that $b^{-1} \in B_k$. Therefore B_k is a subfield of $\mathbb{F}_q$. Hence $g(k,q)=\infty$ if and only if B_k is a proper subfield of $\mathbb{F}_q$. This holds if and only if the multiplicative group G_k of nonzero kth powers in $\mathbb{F}_q$ is a subgroup of $\mathbb{F}_{p^d}^*$ for some proper divisor d of r. Now G_k is cyclic of order $(q-1)/k$ (note that $G_k = \{g^{kj} : j=0,1,\ldots,(q-1)/k-1\}$ with a primitive element g of $\mathbb{F}_q$) and is a subgroup of $\mathbb{F}_{p^d}^*$ if and only if $(q-1)/k$ divides p^d-1, or equivalently if and only if $(q-1)/(p^d-1)$ divides k. □

6.3.2 Addition theorems

Additive number theory studies subsets of $\mathbb{Z}$ (or more generally of abelian groups) and their behavior under addition (or under the binary operation on the abelian group). The principal objects of additive number theory are *sumsets*

$$A+B := \{a+b : a \in A,\ b \in B\},$$

where A and B are nonempty subsets of a given abelian group with the additive notation.

Example 6.3.7 If A and B are nonempty finite sets of real numbers, then we claim that

$$|A+B| \geq |A| + |B| - 1.$$

If we write $A = \{a_1, a_2, \ldots, a_s\}$ with $a_1 < a_2 < \ldots < a_s$ and $B = \{b_1, b_2, \ldots, b_t\}$ with $b_1 < b_2 < \ldots < b_t$, then

$$a_1 + b_1 < a_1 + b_2 < \ldots < a_1 + b_t < a_2 + b_t < \ldots < a_s + b_t,$$

and so at least $s+t-1$ elements of $A+B$ are different. Thus, the claim is established. This lower bound on $|A+B|$ is in general best possible: just take $A = \{0, 1, \ldots, s-1\}$ and $B = \{0, 1, \ldots, t-1\}$ for any $s, t \in \mathbb{N}$.

The analogous result for subsets of a finite field $\mathbb{F}_p$ of prime order p is the Cauchy-Davenport theorem which can be used to prove a general bound on the Waring number $g(k, p)$.

Theorem 6.3.8 (Cauchy-Davenport Theorem) *Let p be a prime number and let A and B be nonempty subsets of $\mathbb{F}_p$. Then*

$$|A+B| \geq \min(|A| + |B| - 1, p).$$

Proof. We present the proof of Alon, Nathanson, and Ruzsa [4]. First we deal with the case where $|A| + |B| \geq p+1$. Then for every $c \in \mathbb{F}_p$ there is an element $a \in A \cap (c - B)$, and thus $A + B = \mathbb{F}_p$.

Now we may assume that $|A| + |B| \leq p$. We consider the space $\mathcal{F}$ of all maps $f : A \times B \to \mathbb{F}_p$ which is a vector space over $\mathbb{F}_p$ of dimension $|A||B|$. Each such map can be identified with a polynomial

$$f(x, y) = \sum_{i=0}^{|A|-1} \sum_{j=0}^{|B|-1} a_{ij} x^i y^j$$

in the variables x and y with coefficients $a_{ij} \in \mathbb{F}_p$, and

$$\mathcal{B} = \{x^i y^j : 0 \leq i \leq |A| - 1,\ 0 \leq j \leq |B| - 1\}$$

is a basis of $\mathcal{F}$ (here you may want to refer to [4, Lemma 2.2]). Put $\mathcal{S} = \mathcal{B} \setminus \{x^{|A|-1} y^{|B|-1}\}$ and note that $x^{|A|-1} y^{|B|-1}$ is not a linear combination over $\mathbb{F}_p$ of elements of $\mathcal{S}$. However, all monomials $x^i y^j$ (as maps from $A \times B$ to $\mathbb{F}_p$) with $0 \leq i < |A| - 1$ and $j \geq |B|$, or $i \geq |A|$ and $0 \leq j < |B| - 1$, are linear combinations over $\mathbb{F}_p$ of elements of $\mathcal{S}$ since x^k, $k \geq |A|$, is a linear combination over $\mathbb{F}_p$ of $\{x^i : 0 \leq i < |A|\}$ and y^k, $k \geq |B|$, is a linear combination over $\mathbb{F}_p$ of $\{y^j : 0 \leq j < |B|\}$.

Now suppose that $|A + B| \leq |A| + |B| - 2$. Then there is a set $C \subset \mathbb{F}_p$ of cardinality $|C| = |A| + |B| - 2$ with $A + B \subseteq C$. We consider the function

$$f(x, y) = \prod_{c \in C} (x + y - c) = \binom{|A| + |B| - 2}{|A| - 1} x^{|A|-1} y^{|B|-1} + \cdots ,$$

which vanishes on $A \times B$. However,

$$\binom{|A| + |B| - 2}{|A| - 1} = \frac{(|A| + |B| - 2)!}{(|A| - 1)!(|B| - 1)!}$$

is not divisible by p (since the factors of the numerator are smaller than p) and we get a contradiction. □

Remark 6.3.9 Here is an elementary alternative proof of the Cauchy-Davenport theorem. Let $|A| = s$, say $A = \{a_1, \ldots, a_s\}$, and $|B| = t$, say $B = \{b_1, \ldots, b_t\}$. We proceed by induction on t. The case $t = 1$ is trivial, and so we take $t \geq 2$. If we put $C = A + B$, then the case $|C| = p$ is trivial, so we can assume $|C| < p$. For $n = 0, 1, \ldots, p - 1$, the elements $a_1 + b_1 + n(b_t - b_1)$ run through $\mathbb{F}_p$ (note that $b_t \neq b_1$), and for $n = 0$ and $n = 1$ we get elements in C. Since $|C| < p$, there exists a least $n_0 \in \mathbb{N}$ with $a_1 + b_1 + n_0(b_t - b_1) \notin C$. Then $a_1 + b_1 + (n_0 - 1)(b_t - b_1) \in C$. With $d := a_1 + b_1 + n_0(b_t - b_1) + b_1$ we obtain $d - b_1 \notin C$ and $d - b_t \in C$. We arrange the elements $b_1, \ldots, b_t$ such that $d - b_i \notin C$ for $1 \leq i \leq r$ and $d - b_j \in C$ for $r < j \leq t$. Clearly $1 \leq r \leq t - 1$. Now we consider the sumset

$$C' := \{a_h + b_i : 1 \leq h \leq s,\ 1 \leq i \leq r\} \subseteq C.$$

Then $d - b_j \in C$ for $r < j \leq t$, but $d - b_j \notin C'$ for $r < j \leq t$, for if we had $d - b_j = a_h + b_i$ for some $1 \leq h \leq s$ and $1 \leq i \leq r$, then we get the contradiction $d - b_i = a_h + b_j \in C$. Therefore $|C'| \leq |C| - (t - r)$. The induction hypothesis yields $|C'| \geq s + r - 1$, and so $|C| \geq |C'| + t - r \geq s + t - 1$ as desired.

Basically the same example as in Example 6.3.7 shows that the Cauchy-Davenport theorem is in general best possible. Now we extend the Cauchy-Davenport theorem in some fashion to arbitrary finite fields. First we have to characterize the binomial coefficients that are divisible by a prime number p via the following congruence of Lucas. We use the standard convention $\binom{m}{n} = 0$ for integers $m, n \geq 0$ with $m < n$.

Lemma 6.3.10 (Lucas Congruence) *If m and n are nonnegative integers and p is a prime number, then*

$$\binom{m}{n} \equiv \prod_{i=0}^{k} \binom{m_i}{n_i} \pmod{p},$$

where

$$m = m_k p^k + m_{k-1}p^{k-1} + \cdots + m_1 p + m_0, \quad 0 \le m_0, \ldots, m_k < p,$$

and

$$n = n_k p^k + n_{k-1}p^{k-1} + \cdots + n_1 p + n_0, \quad 0 \le n_0, \ldots, n_k < p,$$

are the digit expansions in base p of m and n, respectively.

Proof. A computation in the polynomial ring $\mathbb{F}_p[x]$ shows that

$$\begin{aligned}\sum_{n=0}^{m} \binom{m}{n} x^n = (1+x)^m &= \prod_{i=0}^{k} \left((1+x)^{p^i}\right)^{m_i} \\ &= \prod_{i=0}^{k} \left(1 + x^{p^i}\right)^{m_i} = \prod_{i=0}^{k} \Big(\sum_{n_i=0}^{m_i} \binom{m_i}{n_i} x^{n_i p^i} \Big) \\ &= \prod_{i=0}^{k} \Big(\sum_{n_i=0}^{p-1} \binom{m_i}{n_i} x^{n_i p^i} \Big) = \sum_{n=0}^{m} \Big(\prod_{i=0}^{k} \binom{m_i}{n_i} \Big) x^n,\end{aligned}$$

and the result follows by comparing the coefficients of x^n for $n = 0, 1, \ldots, m$. □

Remark 6.3.11 We see from Lemma 6.3.10 that $\binom{m}{n} \equiv 0 \pmod{p}$ if and only if $n_i > m_i$ for some $i = 0, 1, \ldots, k$.

Corollary 6.3.12 *Let q be a power of the prime number p and let A and B be nonempty subsets of the finite field $\mathbb{F}_q$. Then*

$$|A + B| \ge \min(|A| + |B| - q/p, q).$$

Proof. The case $q = p$ is the Cauchy-Davenport theorem, and so we can take $q > p$. If $|A| + |B| \ge q + 1$, then for every $c \in \mathbb{F}_q$ there is an element $a \in A \cap (c - B)$, and thus $A + B = \mathbb{F}_q$.

Now we consider the case where $|A| + |B| \le q$. We can also assume that $|A| > q/p$ and $|B| > q/p$. Let $0 \le s < q/p$ be defined by $|A| - 1 \equiv s \pmod{q/p}$. For any subset A' of A of size $|A| - s$, the Lucas congruence yields $\binom{|A'|+|B|-2}{|A'|-1} \not\equiv 0 \pmod{p}$. As in the proof of Theorem 6.3.8 we get

$$|A' + B| \ge \min(|A'| + |B| - 1, q)$$

and the result follows. □

Now we prove a general bound on $g(k, q)$ which is tight for the examples in Example 6.3.5, but rather weak for most k.

Theorem 6.3.13 *Let q be a prime power and let k be a positive divisor of $q-1$. If $g(k,q)<\infty$, then $g(k,q)\leq k$.*

Proof. In the proof of Theorem 6.3.6 we noted that there are exactly $(q-1)/k$ different nonzero kth powers in $\mathbb{F}_q$. Now we write

$$A_s=\{a_1^k+\cdots+a_s^k : a_1,\ldots,a_s\in\mathbb{F}_q\}$$

for all $s\in\mathbb{N}$. Then either $A_s=\mathbb{F}_q$ or there is an element $b\in A_{s+1}\setminus A_s$. In the latter case we infer that $cb\in A_{s+1}\setminus A_s$ for all $c\in A_1\setminus\{0\}$, and thus

$$|A_{s+1}|\geq|A_s|+|A_1|-1=|A_s|+\frac{q-1}{k}. \tag{6.12}$$

We observe that if q is a prime number, then (6.12) follows directly from the Cauchy-Davenport theorem. By induction we get $|A_s|\geq\min(s\frac{q-1}{k}+1,q)$ for all $s\in\mathbb{N}$, and thus $A_k=\mathbb{F}_q$. □

Further variants and extensions of the Cauchy-Davenport theorem can be found in the book of Nathanson [121] which is also a rich source of information on additive number theory in general.

6.3.3 Sum-product theorems

Sum-product theorems have become a powerful tool for dealing with Waring's problem for finite fields. Roughly speaking, a sum-product theorem for nonempty subsets A of a finite field $\mathbb{F}_q$ says that either the *productset* $A\cdot A=\{ab\in\mathbb{F}_q : a,b\in A\}$ or the sumset $A+A=\{a+b\in\mathbb{F}_q : a,b\in A\}$ is essentially larger than A, provided that there is room to grow, that is, $|A|$ is of smaller order of magnitude than q. We will prove such a sum-product theorem due to Garaev and explain how to derive bounds on $g(k,q)$ from it. Moreover, we will use a sum-product theorem of Glibichuk and Rudnev, where we do not include the proof, to deduce an even stronger result. First we establish the following extension of a result of Garaev [52, Theorem 1].

Theorem 6.3.14 *If A, B, and C are nonempty subsets of the finite field $\mathbb{F}_q$ with $0\notin B$, then*

$$|A\cdot B||A+C|\geq\frac{3-\sqrt{5}}{2}\min\Big(q|A|,\frac{|A|^2|B||C|}{q}\Big).$$

Proof. Let N be the number of solutions of the equation

$$sb^{-1}+c=t,\qquad b\in B,\ c\in C,\ s\in A\cdot B,\ t\in A+C.$$

Each ordered triple $(a,b,c)\in A\times B\times C$ produces a solution of this equation with $s=ab$, $t=a+c$, and different ordered triples give rise to different solutions. Therefore

$$N\geq|A||B||C|. \tag{6.13}$$

If χ is a nontrivial additive character of $\mathbb{F}_q$, then by the orthogonality relation (1.9) we obtain

$$N = \frac{1}{q}\sum_{r\in\mathbb{F}_q}\sum_{b\in B,\, c\in C,\, s\in A\cdot B,\, t\in A+C}\chi(r(sb^{-1}+c-t)).$$

Separating the contribution of $r=0$, we get

$$N \le \frac{|B||C||A\cdot B||A+C|}{q} + \frac{1}{q}\sum_{r\in\mathbb{F}_q^*}\Big|\sum_{b\in B,\, s\in A\cdot B}\chi(rsb^{-1})\Big|\Big|\sum_{c\in C}\chi(rc)\Big|\Big|\sum_{t\in A+C}\chi(rt)\Big|.$$

Next we claim that

$$\Big|\sum_{d\in D,\, e\in E}\chi(de)\Big| \le \sqrt{q|D||E|}$$

for all nonempty subsets D and E of $\mathbb{F}_q$. Indeed, the Cauchy-Schwarz inequality and the orthogonality relation (1.9) imply that

$$\begin{aligned}\Big|\sum_{d\in D,\, e\in E}\chi(de)\Big|^2 &\le \Big(\sum_{d\in D}1\cdot\Big|\sum_{e\in E}\chi(de)\Big|\Big)^2 \le |D|\sum_{d\in D}\Big|\sum_{e\in E}\chi(de)\Big|^2\\ &\le |D|\sum_{r\in\mathbb{F}_q}\Big|\sum_{e\in E}\chi(re)\Big|^2 = |D|\sum_{e_1,e_2\in E}\sum_{r\in\mathbb{F}_q}\chi(r(e_1-e_2)) = q|D||E|,\end{aligned}$$

and the claimed inequality follows.

We obtain

$$\begin{aligned}N \le\ &\frac{|B||C||A\cdot B||A+C|}{q}\\ &+\frac{\sqrt{q|B||A\cdot B|}}{q}\Big(\sum_{r\in\mathbb{F}_q}\Big|\sum_{c\in C}\chi(rc)\Big|^2\Big)^{1/2}\Big(\sum_{r\in\mathbb{F}_q}\Big|\sum_{t\in A+C}\chi(rt)\Big|^2\Big)^{1/2}.\end{aligned}$$

Again by (1.9), we get

$$\sum_{r\in\mathbb{F}_q}\Big|\sum_{c\in C}\chi(rc)\Big|^2 = q|C| \quad\text{and}\quad \sum_{r\in\mathbb{F}_q}\Big|\sum_{t\in A+C}\chi(rt)\Big|^2 = q|A+C|,$$

which together with (6.13) yields the bound

$$|A||B||C| \le \frac{|B||C||A\cdot B||A+C|}{q} + \sqrt{q|B||C||A\cdot B||A+C|}.$$

Simple algebraic manipulations lead to the inequality

$$|A\cdot B||A+C| \ge \left(\sqrt{\frac{q^3}{4|B||C|}+q|A|}-\sqrt{\frac{q^3}{4|B||C|}}\right)^2. \tag{6.14}$$

If we put $u = q^2/(4|A||B||C|)$, then we can write (6.14) as

$$|A\cdot B||A+C| \ge q|A|(\sqrt{u+1}-\sqrt{u})^2 = \frac{q|A|}{(\sqrt{u+1}+\sqrt{u})^2}.$$

If $u \leq \frac{1}{4}$, then $\sqrt{u+1}+\sqrt{u} \leq (\sqrt{5}+1)/2$, and so

$$|A \cdot B||A+C| \geq \frac{4q|A|}{(\sqrt{5}+1)^2} = \frac{3-\sqrt{5}}{2}q|A|.$$

If $u > \frac{1}{4}$, then $\sqrt{1+u^{-1}}+1 < \sqrt{5}+1$, and so $\sqrt{u+1}+\sqrt{u} < (\sqrt{5}+1)\sqrt{u}$. It follows that

$$|A \cdot B||A+C| > \frac{q|A|}{(\sqrt{5}+1)^2 u} = \frac{3-\sqrt{5}}{2} \cdot \frac{|A|^2|B||C|}{q}.$$

Therefore the desired lower bound holds in all cases. □

Example 6.3.15 Let k be a positive divisor of $q-1$, let $A = C = \{a^k : a \in \mathbb{F}_q\}$ be the set of kth powers in $\mathbb{F}_q$, and let $B = A \setminus \{0\}$. Then $|A| = \frac{q-1}{k}+1$ and $A \cdot B = A$, and thus

$$|A+A| \geq \frac{3-\sqrt{5}}{2} \min\Big(q, \frac{|A|^2|B|}{q}\Big) \geq \frac{3-\sqrt{5}}{2} \min\Big(q, \frac{(q-1)^2}{k^3}\Big)$$

by Theorem 6.3.14. On the other hand, (6.12) yields $|A+A| \geq \min\big(2\frac{q-1}{k}+1, q\big) = \frac{2(q-1)}{k}+1$ for $k \geq 2$, which is a smaller bound if $8 \leq k \leq \frac{1}{3}(q-1)^{1/2}$.

Example 6.3.16 Let $q = p$ be a prime number and let H be an integer with $1 \leq H \leq (p+1)/2$. Then for $A = B = C = \{1, \ldots, H\} \subseteq \mathbb{F}_p$ we get $|A+A| = 2H-1 < 2|A|$, and thus

$$|A \cdot A| > \frac{3-\sqrt{5}}{4} \min\Big(p, \frac{H^3}{p}\Big)$$

by Theorem 6.3.14.

Corollary 6.3.17 *For a positive integer s, put $sA = \{a_1+\cdots+a_s : a_1, \ldots, a_s \in A\}$ and $A^s = \{a_1 \cdots a_s : a_1, \ldots, a_s \in A\}$ for every nonempty set $A \subseteq \mathbb{F}_q^*$. Then*

$$|A^s| \cdot |sA| \geq \min\Big(\frac{3-\sqrt{5}}{2}q|A|, \Big(\frac{3-\sqrt{5}}{2}\Big)^s \frac{|A|^{2s}}{q^{s-1}}\Big).$$

Proof. The result follows inductively from Theorem 6.3.14 with $B = A^{s-1}$ and $C = (s-1)A$. □

Example 6.3.18 For a positive divisor k of $q-1$, let $A = \{a^k : a \in \mathbb{F}_q^*\}$ be the set of kth powers in $\mathbb{F}_q^*$. Corollary 6.3.17 implies that

$$|sA| \geq \min\Big(\frac{3-\sqrt{5}}{2}q, \Big(\frac{3-\sqrt{5}}{2}\Big)^s \frac{(q-1)^{2s-1}}{k^{2s-1}q^{s-1}}\Big) \qquad \text{for all } s \in \mathbb{N}.$$

In particular, we have $|sA| \geq \frac{3-\sqrt{5}}{2}q$ if $k \leq c_s q^{(s-1)/(2s-1)}$ with an explicit constant $c_s > 0$ depending only on s. Now applying Corollary 6.3.12 twice, we obtain

$$|3sA| \geq \min\Big(\Big(\frac{3(3-\sqrt{5})}{2} - \frac{2}{p}\Big)q, q\Big) = q \qquad \text{if } k \leq c_s q^{(s-1)/(2s-1)},$$

provided that q is a power of a prime number $p \geq 17$. This yields $g(k,q) \leq 3s$ under the stated conditions on k and q.

Now we present a very strong result which is based on an elegant sum-product theorem of Glibichuk and Rudnev [56].

Theorem 6.3.19 *If k is a positive divisor of $q-1$ with $k \leq \sqrt{q/2}$, then $g(k,q) \leq 8$.*

Proof. By [56, Theorem 6], if $A, B \subseteq \mathbb{F}_q$ with $|A||B| \geq 2q$, then

$$8AB = \{a_1b_1 + \cdots + a_8b_8 : a_1, \ldots, a_8 \in A,\ b_1, \ldots, b_8 \in B\} = \mathbb{F}_q.$$

The result follows by applying this to $A = B = \{a^k : a \in \mathbb{F}_q\}$. □

Remark 6.3.20 For positive divisors k of $q-1$ with $k \leq q^{3/7}$, we can improve on Theorem 6.3.19 by using a well-known result on the number of solutions of diagonal equations obtained via bounds on Jacobi sums (see Exercises 1.36 and 1.37 for the simplest Jacobi sums). For $s \in \mathbb{N}$ and $b \in \mathbb{F}_q^*$, let $N_s(b)$ denote the number of solutions $(c_1, \ldots, c_s) \in \mathbb{F}_q^s$ of

$$c_1^k + \cdots + c_s^k = b.$$

Then [102, Theorem 6.37] yields

$$|N_s(b) - q^{s-1}| < k^s q^{(s-1)/2},$$

and thus

$$N_s(b) > q^{s-1} - k^s q^{(s-1)/2} \geq 0 \qquad \text{for } k \leq q^{1/2-1/(2s)}.$$

Therefore

$$g(k,q) \leq s \qquad \text{if } k \leq q^{1/2-1/(2s)}.$$

This is a trivial result for $s = 1$ and an improvement on Theorem 6.3.19 for $2 \leq s \leq 7$ and the corresponding range for k.

Remark 6.3.21 A *Waring graph* is a directed graph whose vertex set is $\mathbb{F}_q$ and where there is an edge leading from vertex $a \in \mathbb{F}_q$ to vertex $b \in \mathbb{F}_q$ precisely if the difference $b - a$ is a kth power in $\mathbb{F}_q$, with k being a fixed positive divisor of $q-1$. If the Waring number $g(k,q)$ is finite, then $g(k,q)$ can be considered the *diameter* of the Waring graph, that is, the least positive integer r such that any two distinct vertices can be joined by a path consisting of at most r edges. If for suitable k the Waring number $g(k,q)$ is small, then the Waring graph has relatively few edges, but a small diameter. Such graphs are important for computer networks.

There is a profusion of further results on the Waring number $g(k,q)$, and we refer to the survey article [20] for more details.

6.3.4 Covering codes

We like codes so much that we return to them again and again. Now we are covering an aspect that we have not yet covered at all, namely covering codes, which surprisingly are connected with Waring numbers (see Theorem 6.3.29 below). Chapter 3 was devoted to coding theory, and we adhere to the terminology and notation established there.

A *covering code* is a code, that is, a nonempty subset of a Hamming space $(\mathbb{F}_q^n, d)$, with the property that every element of the Hamming space is within a fixed (in the interesting cases small) Hamming distance of some codeword. The standard monograph on covering codes is [26]. Covering codes have many applications including football pool problems and speech coding. Whereas the minimum distance $d(C)$ is the most important quality measure for an error-correcting code C, the main quality measure for a covering code is the covering radius.

Definition 6.3.22 The *covering radius* $\rho(C)$ of a code $C \subseteq \mathbb{F}_q^n$ is

$$\rho(C) = \max_{\mathbf{v} \in \mathbb{F}_q^n} \min_{\mathbf{c} \in C} d(\mathbf{v}, \mathbf{c}).$$

Remark 6.3.23 The football pool problem is based on football betting (or in some countries called soccer betting) where the aim is to correctly predict at least r (with $1 \leq r \leq n$) results of n football matches, that is, home win, draw, or away win, with K bets. Thus, we need a ternary code C of length n and size K with covering radius $\rho(C) \leq n - r$ in order to guarantee that at least one of K bets predicts at least r results correctly.

Remark 6.3.24 In speech coding, S points have to be placed "uniformly" on the surface of a sphere in the n-dimensional Euclidean space. For $S \leq 2^n$, an approximate solution can be obtained by taking the words of a binary code of size S and length n with small covering radius.

There is a nice relationship between the covering radius $\rho(C)$ and the minimum distance $d(C)$ of a code C, and as a bonus we get another characterization of perfect codes (see Definition 3.4.8 for the concept of a perfect code).

Proposition 6.3.25 *A code C is perfect if and only if $d(C) = 2\rho(C)+1$. In general, the bound $d(C) \leq 2\rho(C) + 1$ holds whenever $|C| \geq 2$.*

Proof. Put $d = d(C)$. Then C is perfect if and only if

$$\bigcup_{\mathbf{c} \in C} B(\mathbf{c}, \lfloor (d-1)/2 \rfloor) = \mathbb{F}_q^n,$$

that is, for every $\mathbf{v} \in \mathbb{F}_q^n$ there is exactly one $\mathbf{c} \in C$ with $\mathbf{v}$ in the ball $B(\mathbf{c}, \lfloor (d-1)/2 \rfloor)$ (compare with the proof of Theorem 3.4.6). This is possible only if d is odd. The maximum distance of $\mathbf{v} \in \mathbb{F}_q^n$ to C is $(d-1)/2$, and thus $\rho(C) = (d-1)/2$.

If the code C with $|C| \geq 2$ is not perfect, then we get the proper inclusion

$$\bigcup_{\mathbf{c} \in C} B(\mathbf{c}, \lfloor (d-1)/2 \rfloor) \subset \mathbb{F}_q^n.$$

Thus, there exists a word $\mathbf{u} \in \mathbb{F}_q^n$ that is not contained in any of the balls $B(\mathbf{c}, \lfloor (d-1)/2 \rfloor)$ with $\mathbf{c} \in C$. In other words, $d(\mathbf{u}, \mathbf{c}) \geq \lfloor (d-1)/2 \rfloor + 1$ for all $\mathbf{c} \in C$, and so $\rho(C) \geq \lfloor (d-1)/2 \rfloor + 1 > (d-1)/2$. □

Example 6.3.26 Let

$$C = \{(0, \dots, 0),\ (1, \dots, 1)\} \subseteq \mathbb{F}_2^n$$

be the binary repetition code of length $n \geq 2$. Then $\rho(C) = \lfloor n/2 \rfloor$ and $d(C) = n$.

Example 6.3.27 Let us consider the ternary Golay code G_{11} introduced in Definition 3.5.26. By Theorem 3.5.28, G_{11} is a perfect linear $[11, 6, 5]$ code over $\mathbb{F}_3$. Therefore $\rho(G_{11}) = 2$ by Proposition 6.3.25. Because of its small covering radius, G_{11} is great for football betting (see Remark 6.3.23). In many football pools, the results of 12 football matches have to be predicted every week and there are payouts if at least 10 matches are predicted correctly. Now suppose that among the 12 matches there is one "bank", that is, a match for which you "know" the outcome in your guts. For instance, in the current German Bundesliga a home match by Bayern Munich is a "bank" on Bayern. Then there are only 11 matches left on which you have to bet, and getting at least nine out of these right will earn a payout. Thus, you have the situation in Remark 6.3.23 with the parameters $n = 11$ and $r = 9$. If you are an astute football pool enthusiast, then you place $3^6 = 729$ bets corresponding to the codewords of G_{11} and you will thus hit the jackpot since the condition $\rho(G_{11}) \leq n - r = 2$ is satisfied. This lucrative piece of advice by itself is already worth the price of this book. The only hitch is that you should have enough spare money for 729 bets. It is a truly astounding historical fact that in the context of football betting, the sophisticated code G_{11} was already discovered in 1947 (that is, before Golay's paper [57]) by the Finnish football pool specialist Juhani Virtakallio who published the construction in a Finnish football pool magazine (see [26, Section 15.3] for more details).

For the proof of the following theorem, it is useful to have an alternative description of the covering radius of a linear code at hand.

Lemma 6.3.28 *Let C be a nontrivial linear $[n,k]$ code over $\mathbb{F}_q$ and let H be a parity-check matrix of C. Then the covering radius $\rho(C)$ is the least positive integer r such that every vector in $\mathbb{F}_q^{n-k}$ is a linear combination over $\mathbb{F}_q$ of at most r column vectors of H.*

Proof. Every $\mathbf{u} \in \mathbb{F}_q^{n-k}$ can be written as $\mathbf{u} = \mathbf{v}H^\top$ for some $\mathbf{v} \in \mathbb{F}_q^n$. Let $\mathbf{b} \in C$ be a codeword with

$$d(\mathbf{v},\mathbf{b}) = \min_{\mathbf{c}\in C} d(\mathbf{v},\mathbf{c}) = r \le \rho(C)$$

and let $v_{i_1},\dots,v_{i_r}$ denote the coordinates of $\mathbf{v}$ that differ from the corresponding coordinates of $\mathbf{b}$. Since $\mathbf{b}H^\top = \mathbf{0}$ by Theorem 3.2.37, we get

$$\mathbf{u} = \mathbf{v}H^\top = (\mathbf{v}-\mathbf{b})H^\top = (v_{i_1} - b_{i_1})\mathbf{s}_{i_1} + \cdots + (v_{i_r} - b_{i_r})\mathbf{s}_{i_r},$$

where $\mathbf{s}_i$ denotes the ith column vector of H and b_i the ith coordinate of $\mathbf{b}$. Hence each $\mathbf{u} \in \mathbb{F}_q^{n-k}$ is a linear combination over $\mathbb{F}_q$ of at most $\rho(C)$ column vectors of H.

Conversely, suppose that every vector in $\mathbb{F}_q^{n-k}$ is a linear combination over $\mathbb{F}_q$ of at most r column vectors of H. Let $\mathbf{v} \in \mathbb{F}_q^n$ be arbitrary and put $\mathbf{u} = \mathbf{v}H^\top \in \mathbb{F}_q^{n-k}$. Then by assumption $\mathbf{u} = \mathbf{x}H^\top$ for some $\mathbf{x} \in \mathbb{F}_q^n$ with Hamming weight $w(\mathbf{x}) \le r$. It follows that $(\mathbf{v}-\mathbf{x})H^\top = \mathbf{0}$, and so $\mathbf{v}-\mathbf{x} = \mathbf{b}$ for some $\mathbf{b} \in C$ by Theorem 3.2.37. Therefore

$$\min_{\mathbf{c}\in C} d(\mathbf{v},\mathbf{c}) = \min_{\mathbf{c}\in C} w(\mathbf{v}-\mathbf{c}) \le w(\mathbf{v}-\mathbf{b}) = w(\mathbf{x}) \le r,$$

and so $\rho(C) \le r$. □

Theorem 6.3.29 *Let q be a prime power, let $n \ge 2$ be an integer with $\gcd(n,q) = 1$, let h be the multiplicative order of q modulo n, and let $\alpha \in \mathbb{F}_{q^h}$ be a primitive nth root of unity. Let $m(x) \in \mathbb{F}_q[x]$ be the minimal polynomial of α over $\mathbb{F}_q$. Then the cyclic code $C \subseteq \mathbb{F}_q^n$ with generator polynomial $m(x)$ satisfies*

$$\rho(C) \le g((q^h-1)/n, q^h),$$

with equality for $q = 2$.

Proof. We refer to Subsection 3.3.5 for the general construction of cyclic codes from roots. For $\mathbf{v} = (v_0, v_1, \dots, v_{n-1}) \in \mathbb{F}_q^n$, Theorem 3.3.30 shows that $\mathbf{v} \in C$ if and only if $\sum_{j=0}^{n-1} v_j\alpha^j = 0$. This means that the matrix

$$H = (1 \ \alpha \ \alpha^2 \ \dots \ \alpha^{n-1})$$

can be viewed as a parity-check matrix of C when each α^j, $j = 0,1,\dots,n-1$, is replaced by its coordinate vector (written as a column vector) relative to a fixed ordered basis of $\mathbb{F}_{q^h}$ over $\mathbb{F}_q$. Furthermore, we note that the cyclic subgroup of $\mathbb{F}_{q^h}^*$ generated by α agrees with the cyclic subgroup of $\mathbb{F}_{q^h}^*$ consisting of the nonzero kth powers with $k = (q^h-1)/n$. The rest follows from Lemma 6.3.28 and the definition of the Waring number $g(k,q)$ in Definition 6.3.3. □

6.4 Hadamard matrices and applications

6.4.1 Basic constructions

Hadamard matrices are fascinating combinatorial objects allowing several applications including error-correcting codes, mobile communication, radar/sonar, and cryptography.

Definition 6.4.1 A *Hadamard matrix* H of order $n \geq 1$ is an $n \times n$ matrix over $\mathbb{R}$ with entries 1 or -1 such that $HH^\top = nE_n$, where E_n is the $n \times n$ identity matrix over $\mathbb{R}$.

Example 6.4.2 Examples of Hadamard matrices of the three least possible orders are (1), $\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, and

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

Remark 6.4.3 If you are familiar with determinants, then you understand that

$$(\det(H))^2 = \det(H)\det(H^\top) = \det(HH^\top) = \det(nE_n) = n^n$$

for every Hadamard matrix H of order n, and so $\det(H) = \pm n^{n/2}$. This is remarkable in the light of a classical inequality of Jacques Hadamard (1865–1963) which says that $|\det(M)| \leq n^{n/2}$ for every $n \times n$ matrix M over $\mathbb{R}$ with all entries of absolute value at most 1. Therefore Hadamard matrices are optimal in this family of matrices M in the sense that they meet this bound. By the way, Hadamard is famous in number theory since he is one of the two mathematicians (the other one is de la Vallée-Poussin) who first proved the prime number theorem, according to which the number $\pi(u)$ of prime numbers not exceeding $u \in \mathbb{R}$ is asymptotically equal to $u/\log u$ as $u \to \infty$.

Lemma 6.4.4 *If H is a Hadamard matrix of order $n \geq 3$, then n is divisible by* 4.

Proof. This is basically a one-liner. Let $H = (h_{ij})_{1 \leq i,j \leq n}$ with $n \geq 3$. From $HH^\top = nE_n$ we obtain

$$\sum_{j=1}^{n} (h_{1j} + h_{2j})(h_{1j} + h_{3j}) = \sum_{j=1}^{n} h_{1j}^2 = n.$$

Every term in the first sum is either 0 or 4, hence the result follows. □

The *Hadamard matrix conjecture* claims that for every $n \in \mathbb{N}$ divisible by 4 there is a Hadamard matrix of order n. Currently, the smallest open cases are $n = 668$ and $n = 716$.

Lemma 6.4.5 *If H is a Hadamard matrix of order n, then*

$$\begin{pmatrix} H & H \\ H & -H \end{pmatrix}$$

is a Hadamard matrix of order $2n$.

Proof. The matrix computation

$$\begin{aligned}\begin{pmatrix} H & H \\ H & -H \end{pmatrix}\begin{pmatrix} H & H \\ H & -H \end{pmatrix}^{\top} &= \begin{pmatrix} HH^{\top}+HH^{\top} & HH^{\top}-HH^{\top} \\ HH^{\top}-HH^{\top} & HH^{\top}+HH^{\top} \end{pmatrix} \\ &= \begin{pmatrix} 2nE_n & 0 \\ 0 & 2nE_n \end{pmatrix} = 2nE_{2n}\end{aligned}$$

shows the result. □

The matrices obtained by the iterative construction $S_0 = (1)$ and

$$S_t = \begin{pmatrix} S_{t-1} & S_{t-1} \\ S_{t-1} & -S_{t-1} \end{pmatrix} \qquad \text{for } t = 1, 2, \ldots$$

are called *Sylvester matrices*. They are all Hadamard matrices by Lemma 6.4.5. Sylvester matrices solve the existence problem for Hadamard matrices for all orders $n = 2^t$, $t = 0, 1, \ldots$. Here is a nice explicit formula for Sylvester matrices. Write any two integers $0 \le i, j \le 2^t - 1$ in their unique binary representation

$$i = i_0 + 2i_1 + 4i_2 + \cdots + 2^{t-1}i_{t-1}, \quad i_0, i_1, i_2, \ldots, i_{t-1} \in \{0, 1\},$$

$$j = j_0 + 2j_1 + 4j_2 + \cdots + 2^{t-1}j_{t-1}, \quad j_0, j_1, j_2, \ldots, j_{t-1} \in \{0, 1\},$$

and then put

$$< i, j >:= i_0j_0 + i_1j_1 + \cdots + i_{t-1}j_{t-1}.$$

Theorem 6.4.6 *Sylvester matrices have the explicit form*

$$S_t = \left((-1)^{<i,j>}\right)_{0 \le i,j < 2^t} \qquad \textit{for all } t \ge 0.$$

Proof. Proceed by induction on t. □

Although Hadamard matrices are real matrices, there is a legendary construction of Hadamard matrices due to Paley [157] which intriguingly enough uses finite fields. Raymond Paley (1907–1933) died in a skiing accident in Banff the same year his paper on orthogonal matrices was published. Standing at his grave near the Banff International Research Station BIRS, the second author wondered whether this should deter him from publishing a big result and then dying or from learning Alpine skiing despite living in Austria for already 15 years.

The basic tool of Payley's construction is the quadratic character η of a finite field of odd order q (see Remark 1.4.53). We use the convention $\eta(b) = 0$ for $b = 0 \in \mathbb{F}_q$. The following simple character sum identity plays a crucial role.

Lemma 6.4.7 *Let q be a power of an odd prime and let η be the quadratic character of $\mathbb{F}_q$. Then*

$$\sum_{c\in\mathbb{F}_q} \eta(c)\eta(c+a) = -1 \qquad \text{for all } a \in \mathbb{F}_q^*.$$

Proof. Simple manipulations show that

$$\begin{aligned}\sum_{c\in\mathbb{F}_q} \eta(c)\eta(c+a) &= \sum_{c\in\mathbb{F}_q^*\setminus\{-a\}} \eta(c^{-1}(c+a)) = \sum_{c\in\mathbb{F}_q^*\setminus\{-a\}} \eta(1+ac^{-1}) \\ &= \sum_{c\in\mathbb{F}_q^*\setminus\{1\}} \eta(c) = -\eta(1) = -1\end{aligned}$$

for all $a \in \mathbb{F}_q^*$. We used the orthogonality relation (1.9) in the penultimate step. □

Theorem 6.4.8 *Let q be a power of an odd prime. If $q \equiv 3 \pmod 4$, then we can construct a Hadamard matrix of order $q+1$. If $q \equiv 1 \pmod 4$, then we can construct a Hadamard matrix of order $2(q+1)$.*

Proof. Let η be the quadratic character of $\mathbb{F}_q$. We set up the $q \times q$ matrix $P = (p_{a,b})_{a,b\in\mathbb{F}_q}$ with $p_{a,b} := \eta(b-a)$. For $q \equiv 3 \pmod 4$ we get a Hadamard matrix H of order $q+1$ by substituting the entries 0 in the main diagonal of P by -1 and appending a row and a column consisting of all 1 entries. The fact that H is indeed a Hadamard matrix is shown by a straightforward computation using (1.9), Lemma 6.4.7, and $\eta(-1) = -1$ for $q \equiv 3 \pmod 4$.

For $q \equiv 1 \pmod 4$ we put

$$J = \begin{pmatrix} 0 & 1\ldots 1 \\ (1\ldots 1)^\top & P \end{pmatrix}$$

and

$$H = \begin{pmatrix} J + E_{q+1} & J - E_{q+1} \\ J - E_{q+1} & -J - E_{q+1} \end{pmatrix}.$$

A somewhat more involved computation shows that H is a Hadamard matrix of order $2(q+1)$. □

The Hadamard matrices in the proof of Theorem 6.4.8 are called *Paley matrices*. Here are examples for the two smallest possible values of q.

Example 6.4.9 Let $q = 3$. Then $\eta(1) = 1$ and $\eta(2) = -1$, hence

$$P = \begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}, \quad H = \begin{pmatrix} \mathbf{-1} & 1 & -1 & \mathbf{1} \\ -1 & \mathbf{-1} & 1 & \mathbf{1} \\ 1 & -1 & \mathbf{-1} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \end{pmatrix},$$

and H is a Hadamard matrix of order 4.

Example 6.4.10 Now let $q = 5$. Then $\eta(1) = \eta(4) = 1$ and $\eta(2) = \eta(3) = -1$. Therefore

$$P = \begin{pmatrix} 0 & 1 & -1 & -1 & 1 \\ 1 & 0 & 1 & -1 & -1 \\ -1 & 1 & 0 & 1 & -1 \\ -1 & -1 & 1 & 0 & 1 \\ 1 & -1 & -1 & 1 & 0 \end{pmatrix}, \quad J = \begin{pmatrix} \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & 0 & 1 & -1 & -1 & 1 \\ \mathbf{1} & 1 & 0 & 1 & -1 & -1 \\ \mathbf{1} & -1 & 1 & 0 & 1 & -1 \\ \mathbf{1} & -1 & -1 & 1 & 0 & 1 \\ \mathbf{1} & 1 & -1 & -1 & 1 & 0 \end{pmatrix}.$$

A Hadamard matrix of order 12 is obtained from J as in the proof of Theorem 6.4.8.

More on Hadamard matrices can be found in the book of Horadam [69], but not all the information there is up-to-date because of the steady progress in this area. A status report up to the year 2010 was given by the same author in [70].

6.4.2 Hadamard codes

Hadamard matrices have applications to various aspects of information theory and digital communication. We start with an application to coding theory. Recall that the entries of Hadamard matrices have only two possible values 1 or -1, and so it is pretty obvious that Hadamard matrices are relevant only to binary codes. We may change -1 to 0 to get the standard alphabet $\mathbb{F}_2$ of a binary code. If an arbitrary (that is, not necessarily linear) binary code C has length n, size M, and minimum distance d, then we express this by saying that C is a binary (n, M, d) code. We introduce an interesting quantity that was not considered in Chapter 3 on coding theory.

Definition 6.4.11 For integers n and d with $1 \leq d \leq n$, let $A(n, d)$ be the largest possible integer M for which there exists a binary (n, M, d) code. A binary (n, M, d) code with $M = A(n, d)$ is called *optimal*.

The determination of $A(n, d)$ is a major problem in coding theory. Unfortunately, only partial information on $A(n, d)$ is available, mainly in the form of lower and upper bounds. Hadamard matrices are instrumental in determining certain values of $A(n, d)$ (see the proof of Theorem 6.4.14). We set off with a simple inequality.

Lemma 6.4.12 *If n and d are integers with $1 \leq d \leq n - 1$, then*

$$A(n, d) \leq 2A(n - 1, d).$$

Proof. Let C be an arbitrary binary (n, M, d) code, and for $a \in \{0, 1\}$ let M_a be the number of codewords in C with last coordinate a. Then by deleting the last coordinate we get $(n - 1, M_a, d)$ codes C_a for $a \in \{0, 1\}$ (we may assume that $M_0 \geq 1$ and $M_1 \geq 1$). Obviously $\max(M_0, M_1) \leq A(n-1, d)$, and thus $M = M_0 + M_1 \leq 2A(n-1, d)$. □

In Theorem 3.4.19 we established the Plotkin bound for linear codes. Here is a version of the Plotkin bound for arbitrary binary codes.

Theorem 6.4.13 (Plotkin Bound) *If n and d are integers with $1 \leq d \leq n$ and $n < 2d$, then*

$$A(n, d) \leq \left\lfloor \frac{2d}{2d-n} \right\rfloor .$$

Proof. Let C be an arbitrary binary (n, M, d) code and put

$$T = \sum_{\mathbf{c}, \mathbf{c}' \in C} d(\mathbf{c}, \mathbf{c}').$$

From the definition of the minimum distance we obtain

$$T \geq M(M-1)d. \tag{6.15}$$

For $i = 1, \ldots, n$ and $a \in \{0, 1\}$, let $N_{i,a}$ be the number of codewords in C with ith coordinate equal to a. Writing $\mathbf{c} = (c_1, \ldots, c_n)$ and $\mathbf{c}' = (c'_1, \ldots, c'_n)$, we get

$$\begin{aligned} T &= \sum_{i=1}^{n} \Big(\sum_{\mathbf{c}, \mathbf{c}' \in C} d(c_i, c'_i) \Big) = \sum_{i=1}^{n} \sum_{a \in \{0,1\}} N_{i,a}(M - N_{i,a}) \\ &= M^2 n - \sum_{i=1}^{n} (N_{i,0}^2 + N_{i,1}^2) \leq M^2 n - \frac{1}{2} \sum_{i=1}^{n} (N_{i,0} + N_{i,1})^2 = M^2 n/2, \end{aligned}$$

where we used the elementary inequality $2(u^2 + v^2) \geq (u+v)^2$ for all $u, v \in \mathbb{R}$ in the penultimate step. Together with (6.15) the result follows. □

Theorem 6.4.14 *If d is a positive integer for which there exists a Hadamard matrix of order $4d$, then*

$$A(4d, 2d) = 8d \quad \textit{and} \quad A(4d-1, 2d) = 4d.$$

Proof. Let H be a Hadamard matrix of order $4d$. The $4d$ rows of H are in $\{-1, 1\}^{4d}$. Since $HH^\top = 4dE_{4d}$, the standard inner product (on $\mathbb{R}^{4d}$) of any two distinct rows of H is 0, and so $2d$ entries of the two rows agree and $2d$ entries differ. Interpreted in this way, the Hamming distance of any two distinct rows of H is $2d$. The same holds true for the matrix $-H$ that we obtain by multiplying all entries of H by -1.

Now let C be the binary code of size $8d$ whose codewords are the rows of H and $-H$, with -1 replaced by 0. Then the distinct codewords in C have (standard) Hamming distance either $2d$ or $4d$, with $2d$ actually appearing. Hence C is a binary $(4d, 8d, 2d)$ code, and so $A(4d, 2d) \geq 8d$. Lemma 6.4.12 implies $2A(4d-1, 2d) \geq A(4d, 2d) \geq 8d$, and by the Plotkin bound in Theorem 6.4.13 we get $A(4d-1, 2d) \leq 4d$ and thus the result. □

The optimal code C in the proof of Theorem 6.4.14 is called a *Hadamard code*. The binary Reed-Muller code $\mathcal{R}(1,5)$ mentioned in Section 3.6 is a Hadamard code defined with a Hadamard matrix of order 32 and it is an optimal binary $(32, 64, 16)$ code..

We sketch an application of Hadamard matrices to wireless communication. *CDMA* (*code division multiple access*) is a technology that allows several transmitters to send information simultaneously over the same channel. It is one of the most widely used standards in mobile phone networks and in the GPS (Global Positioning System). Suppose we have M participants $P_1, \ldots, P_M$ in the network. Just to explain the principle in simple terms, let us say that each participant P_i is assigned a periodic signature sequence with full period

$$x_i(1), \ldots, x_i(n) \qquad \text{for } i = 1, \ldots, M.$$

The signal for the symbol a of participant P_i is the sequence with full period

$$ax_i(1), \ldots, ax_i(n).$$

The signals of different participants have to be distinguishable as much as possible. The distinct rows of a Hadamard matrix H of order $n \geq 2$ differ in exactly $n/2$ entries, by the argument in the proof of Theorem 6.4.14, and so the rows of H (periodically continued) are suitable signatures for at most $M = n$ participants. For example, the industry standard QUALCOMM uses a Hadamard matrix of order 64.

6.4.3 Signal correlation

For radar or sonar, a signal is used to determine distances by comparing the original signal $s(0), s(1), \ldots, s(N-1)$ with its time-delayed (or shifted) signal $s(t), s(t+1), \ldots, s(N-1+t)$. Formally, we can think of these signals as periodically continued sequences $(s(n))_{n=0}^{\infty}$ and $(s(n+t))_{n=0}^{\infty}$ with period length N. It is convenient to say that a sequence is N-periodic if it is periodic with period length (not necessarily the least period length) equal to N. The following concept is important in the context of signal processing.

Definition 6.4.15 The *autocorrelation function* of an N-periodic sequence $\sigma = (s(n))_{n=0}^{\infty}$ of complex numbers is defined by

$$A_\sigma(t) = \sum_{n=0}^{N-1} s(n)\overline{s(n+t)} \qquad \text{for } t = 0, 1, \ldots, N-1,$$

where the bar denotes complex conjugation.

In many practical applications the signal is a binary sequence, and in this case links with number theory arise. We assume without loss of generality that the terms of the binary sequence are 1 or -1. The aim is to construct N-periodic binary sequences σ for which $|A_\sigma(t)|$ is small for $1 \leq t \leq N-1$. Note that we always have $A_\sigma(0) = N$.

Proposition 6.4.16 *If σ is an N-periodic binary sequence with terms ± 1, then*

$$A_\sigma(t) \equiv N \pmod 4 \qquad \text{for } t = 0, 1, \ldots, N-1.$$

Proof. Let t be fixed. For $j, k = \pm 1$, let $D_{j,k}$ be the number of integers n with $0 \leq n \leq N-1$ and $(s(n), s(n+t)) = (j, k)$. Then

$$A_\sigma(t) = D_{1,1} + D_{-1,-1} - D_{1,-1} - D_{-1,1}.$$

By counting the number of $0 \leq n \leq N-1$ with $s(n) = 1$ in two ways, we obtain

$$D_{1,-1} + D_{1,1} = D_{-1,1} + D_{1,1},$$

and this implies $D_{1,-1} = D_{-1,1}$. Moreover,

$$N = D_{1,1} + D_{-1,-1} + D_{1,-1} + D_{-1,1} = D_{1,1} + D_{-1,-1} + 2D_{1,-1}.$$

Hence

$$A_\sigma(t) = D_{1,1} + D_{-1,-1} - 2D_{1,-1} = N - 4D_{1,-1}$$

and the result follows. □

Example 6.4.17 The 4-periodic sequence σ obtained by the periodic continuation of the signal $1, 1, 1, -1$ has autocorrelation function $A_\sigma(0) = 4$ and $A_\sigma(t) = 0$ for $t = 1, 2, 3$.

Remark 6.4.18 If we write the N shifts of an N-periodic binary sequence σ with $A_\sigma(t) = 0$ for $t = 1, \ldots, N-1$ as rows of a matrix, then we get a *circulant Hadamard matrix*. However, the *circulant Hadamard matrix conjecture* claims that no circulant Hadamard matrix of order $N > 4$ exists. This conjecture has been verified for a large finite range of values of N (see [99]).

A remarkable number-theoretic sequence with small autocorrelation function is the *Legendre sequence* $\lambda = (\ell(n))_{n=0}^{\infty}$ for the odd prime modulus p which is given by $\ell(n) = 1$ if $n \equiv 0 \pmod p$ and $\ell(n) = \left(\frac{n}{p}\right) = \eta(n)$ for $n \not\equiv 0 \pmod p$, where $\left(\frac{n}{p}\right)$ is the Legendre symbol (see Definition 1.2.22) and η is the quadratic character of $\mathbb{F}_p$. It is obvious that λ is a p-periodic binary sequence.

Theorem 6.4.19 *The autocorrelation function $A_\lambda(t)$, $1 \le t \le p-1$, of the Legendre sequence λ for the odd prime modulus p is given by*

$$A_\lambda(t) = \begin{cases} 2\eta(t) - 1 & \text{if } p \equiv 1 \pmod 4, \\ -1 & \text{if } p \equiv 3 \pmod 4. \end{cases}$$

Proof. For $t = 1, \dots, p-1$, we obtain

$$A_\lambda(t) = \eta(t) + \eta(-t) + \sum_{n \in \mathbb{F}_p} \eta(n)\eta(n+t) = \eta(t) + \eta(-t) - 1$$

by Lemma 6.4.7, and it remains to observe that $\eta(-1) = (-1)^{(p-1)/2}$ by Example 1.2.25. □

Remark 6.4.20 For an integer $d \ge 2$, let $\mathcal{B}_d = (a_n)_{n=1}^{\infty}$ be the maximal period sequence over $\mathbb{F}_2$ introduced in (5.29) and Remark 5.4.6. Note that $\mathcal{B}_d$ is periodic with least period length $2^d - 1$. We can easily turn $\mathcal{B}_d$ into a (2^d-1)-periodic binary sequence $\sigma_d = (s(n))_{n=0}^{\infty}$ with terms ±1 by putting

$$s(n) = (-1)^{a_{n+1}} \qquad \text{for } n = 0, 1, \dots .$$

There is only one nontrivial additive character χ of $\mathbb{F}_2$ and it is given by $\chi(c) = (-1)^c$ for $c \in \mathbb{F}_2 = \{0, 1\}$. The autocorrelation function $A_{\sigma_d}(t)$ of σ_d can be computed for $t = 1, \dots, 2^d - 2$ by

$$A_{\sigma_d}(t) = \sum_{n=0}^{2^d-2} s(n)s(n+t) = \sum_{n=0}^{2^d-2} (-1)^{a_{n+1} - a_{n+t+1}} = \sum_{n=1}^{2^d-1} \chi(a_n - a_{n+t}) = -1,$$

where we used Theorem 5.4.8 in the last step. The binary sequences σ_d derived from maximal period sequences over $\mathbb{F}_2$ have thus an extremely small autocorrelation function. In fact, since $A_\sigma(t) \equiv 2^d - 1 \equiv -1 \pmod 4$ for every (2^d-1)-periodic binary sequence σ by Proposition 6.4.16, the values of $A_{\sigma_d}(t)$ for $1 \le t \le 2^d - 2$ are as close to 0 as possible. This optimality property explains why the binary sequences σ_d are highly popular in signal processing.

6.4.4 Hadamard transform and bent functions

Now we turn to connections between Hadamard matrices and cryptography, and in particular block ciphers (see Section 2.2). We start with some basic concepts.

Definition 6.4.21 A *Boolean function* f (of n variables) is a map $f : \mathbb{F}_2^n \to \mathbb{F}_2$. The associated *binary function* F with values $\pm 1 \in \mathbb{Z}$ is defined by

$$F(\mathbf{u}) = (-1)^{f(\mathbf{u})} \qquad \text{for all } \mathbf{u} \in \mathbb{F}_2^n.$$

The Boolean functions that are given by the dot products

$$l_{\mathbf{v}}(\mathbf{u}) = \mathbf{u} \cdot \mathbf{v} = u_1 v_1 + \cdots + u_n v_n \in \mathbb{F}_2$$

with variable $\mathbf{u} = (u_1, \ldots, u_n) \in \mathbb{F}_2^n$ and fixed $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{F}_2^n$ are called *linear*, and they are *affine* if a constant $c \in \mathbb{F}_2$ is added to $l_{\mathbf{v}}(\mathbf{u})$. Boolean functions suitable for cryptography should not be "close" to any affine Boolean function, in a sense that can be made precise (see [160, Section 3.6]).

Definition 6.4.22 The *Hadamard transform* of the binary function F on $\mathbb{F}_2^n$ (or of the corresponding Boolean function f of n variables) is the integer-valued function $\hat{F}$ on $\mathbb{F}_2^n$ given by

$$\hat{F}(\mathbf{u}) = \sum_{\mathbf{v} \in \mathbb{F}_2^n} (-1)^{\mathbf{u} \cdot \mathbf{v}} F(\mathbf{v}) \qquad \text{for all } \mathbf{u} \in \mathbb{F}_2^n.$$

Note the close connection between the coefficients $(-1)^{\mathbf{u} \cdot \mathbf{v}}$ in the Hadamard transform and the entries of the Sylvester matrix S_n in Theorem 6.4.6. The hat symbol was already used in Section 4.3 for Fourier coefficients of periodic functions on $\mathbb{R}^s$. In the present subsection it will be employed only for the Hadamard transform, and so there should not be any danger of confusion.

Remark 6.4.23 We observe that $\hat{F}(\mathbf{u})$ can be written also as

$$\hat{F}(\mathbf{u}) = \sum_{\mathbf{v} \in \mathbb{F}_2^n} (-1)^{\mathbf{u} \cdot \mathbf{v} + f(\mathbf{v})},$$

and so $\hat{F}(\mathbf{u})$ is the difference of the numbers of $\mathbf{v} \in \mathbb{F}_2^n$ for which $\mathbf{u} \cdot \mathbf{v}$ and $f(\mathbf{v})$ are equal or different, respectively. The Hadamard transform provides one way of measuring how well f can be approximated by linear or affine Boolean functions. For Boolean functions suitable for cryptography, the number

$$\max_{\mathbf{u} \in \mathbb{F}_2^n} |\hat{F}(\mathbf{u})|$$

has to be small (see again [160, Section 3.6]).

Proposition 6.4.24 (Parseval Identity) *Every binary function F on $\mathbb{F}_2^n$ satisfies*

$$\sum_{\mathbf{u} \in \mathbb{F}_2^n} \hat{F}(\mathbf{u})^2 = 2^{2n}.$$

Proof. An easy computation yields

$$\begin{aligned} \sum_{\mathbf{u} \in \mathbb{F}_2^n} \hat{F}(u)^2 &= \sum_{\mathbf{u} \in \mathbb{F}_2^n} \Big(\sum_{\mathbf{v} \in \mathbb{F}_2^n} (-1)^{\mathbf{u} \cdot \mathbf{v}} F(\mathbf{v}) \Big)^2 \\ &= \sum_{\mathbf{u} \in \mathbb{F}_2^n} \sum_{\mathbf{v}, \mathbf{w} \in \mathbb{F}_2^n} (-1)^{\mathbf{u} \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{w}} F(\mathbf{v}) F(\mathbf{w}) \\ &= \sum_{\mathbf{v}, \mathbf{w} \in \mathbb{F}_2^n} F(\mathbf{v}) F(\mathbf{w}) \sum_{\mathbf{u} \in \mathbb{F}_2^n} (-1)^{\mathbf{u} \cdot (\mathbf{v} - \mathbf{w})}. \end{aligned}$$

For $\mathbf{v} = \mathbf{w}$ the inner sum is equal to 2^n. For $\mathbf{v} \neq \mathbf{w}$ there exists a vector $\mathbf{x} \in \mathbb{F}_2^n$ with $\mathbf{x} \cdot (\mathbf{v} - \mathbf{w}) = 1$. Then $\chi(\mathbf{u}) = (-1)^{\mathbf{u}\cdot(\mathbf{v}-\mathbf{w})}$ for all $\mathbf{u} \in \mathbb{F}_2^n$ defines a nontrivial character of the finite abelian group $\mathbb{F}_2^n$ under vector addition, and so the orthogonality relation (1.9) implies that the inner sum vanishes. Therefore

$$\sum_{\mathbf{v},\mathbf{w}\in\mathbb{F}_2^n} F(\mathbf{v})F(\mathbf{w}) \sum_{\mathbf{u}\in\mathbb{F}_2^n} (-1)^{\mathbf{u}\cdot(\mathbf{v}-\mathbf{w})} = \sum_{\mathbf{v}\in\mathbb{F}_2^n} F(\mathbf{v})^2 2^n = 2^{2n},$$

and the result follows. □

Corollary 6.4.25 *Every binary function F on $\mathbb{F}_2^n$ satisfies*

$$\max_{\mathbf{u}\in\mathbb{F}_2^n} |\hat{F}(\mathbf{u})| \geq 2^{n/2}.$$

Proof. This is an immediate consequence of Proposition 6.4.24. □

Corollary 6.4.25 imposes a restriction on how small we can make the quantity $\max_{\mathbf{u}\in\mathbb{F}_2^n} |\hat{F}(\mathbf{u})|$ in Remark 6.4.23. If you are ambitious, then you will strive to achieve equality in the lower bound in Corollary 6.4.25. This inflicts a serious limitation, because then the Parseval identity shows that we must have $\hat{F}(\mathbf{u})^2 = 2^n$ for all $\mathbf{u} \in \mathbb{F}_2^n$. These functions F are singled out by the following terminology.

Definition 6.4.26 The binary function F on $\mathbb{F}_2^n$ (or the corresponding Boolean function f of n variables) is called *bent* if

$$|\hat{F}(\mathbf{u})| = 2^{n/2} \qquad \text{for all } \mathbf{u} \in \mathbb{F}_2^n.$$

Since the Hadamard transform is integer-valued, it is obvious that bent functions can exist only if the number n of variables is even. Here is an appealing link between bent functions and Hadamard matrices.

Theorem 6.4.27 *Let f be a Boolean function of n variables and let F be the associated binary function. Then the following three assertions are equivalent:*
(i) *f is bent;*
(ii) *$(2^{-n/2}\hat{F}(\mathbf{u}+\mathbf{v}))_{\mathbf{u},\mathbf{v}\in\mathbb{F}_2^n}$ is a Hadamard matrix of order 2^n;*
(iii) *$(F(\mathbf{u}+\mathbf{v}))_{\mathbf{u},\mathbf{v}\in\mathbb{F}_2^n}$ is a Hadamard matrix of order 2^n.*

Proof. Note that f ist bent if and only if $2^{-n/2}\hat{F}(\mathbf{u}) = \pm 1$ for all $\mathbf{u} \in \mathbb{F}_2^n$. For every nonzero vector $\mathbf{v} \in \mathbb{F}_2^n$ and every binary function F on $\mathbb{F}_2^n$, we get

$$\sum_{\mathbf{u}\in\mathbb{F}_2^n} \hat{F}(\mathbf{u})\hat{F}(\mathbf{u}+\mathbf{v}) = 2^n \sum_{\mathbf{w}\in\mathbb{F}_2^n} (-1)^{\mathbf{v}\cdot\mathbf{w}} F(\mathbf{w})^2 = 0$$

by similar arguments as in the proof of Proposition 6.4.24. Therefore (i) and (ii) are equivalent.

Now assume that (iii) is true, that is,

$$\sum_{\mathbf{u}\in\mathbb{F}_2^n} F(\mathbf{u})F(\mathbf{u}+\mathbf{v}) = 0$$

for all nonzero vectors $\mathbf{v}\in\mathbb{F}_2^n$. Then with $\mathbf{x}=\mathbf{v}+\mathbf{w}$ we get

$$\hat{F}(\mathbf{u})^2 = \sum_{\mathbf{v},\mathbf{w}\in\mathbb{F}_2^n} (-1)^{\mathbf{u}\cdot(\mathbf{v}+\mathbf{w})}F(\mathbf{v})F(\mathbf{w}) = \sum_{\mathbf{x}\in\mathbb{F}_2^n}(-1)^{\mathbf{u}\cdot\mathbf{x}}\sum_{\mathbf{w}\in\mathbb{F}_2^n} F(\mathbf{w})F(\mathbf{w}+\mathbf{x}) = 2^n$$

and f is bent. Conversely, assume that f is bent, and thus (ii) holds. Then also the Boolean function g defined by $(-1)^{g(\mathbf{u})} = 2^{-n/2}\hat{F}(\mathbf{u})$ for all $\mathbf{u}\in\mathbb{F}_2^n$ is bent. Denote by $\hat{G}$ the Hadamard transform of g. Then $F(\mathbf{u}) = 2^{-n/2}\hat{G}(\mathbf{u})$ for all $\mathbf{u}\in\mathbb{F}_2^n$, and so (iii) follows from (ii). □

Example 6.4.28 Let $n=2m$ with $m\in\mathbb{N}$ and consider the Boolean function

$$f(u_1,\dots,u_{2m}) = u_1u_2 + u_3u_4 + \cdots + u_{2m-1}u_{2m}.$$

The value $\hat{F}(\mathbf{u})$ of the Hadamard transform of f at $\mathbf{u}=(u_1,\dots,u_{2m})\in\mathbb{F}_2^{2m}$ is given by

$$\hat{F}(\mathbf{u}) = \sum_{\mathbf{v}\in\mathbb{F}_2^{2m}} (-1)^{\mathbf{u}\cdot\mathbf{v}}(-1)^{f(\mathbf{v})} = \prod_{j=1}^{m}\Big(\sum_{v,w\in\mathbb{F}_2}(-1)^{u_{2j-1}v+u_{2j}w+vw}\Big).$$

By distinguishing the cases $u_{2j-1}=u_{2j}=0$, $u_{2j-1}=u_{2j}=1$, and $u_{2j-1}\neq u_{2j}$, we see that the last double sum has the value ± 2. Therefore $\hat{F}(\mathbf{u}) = \pm 2^m$, and so f is bent. An entire cottage industry is devoted to the construction of bent functions; we refer to [78] for a recent survey.

There are several cryptographic quality measures for Boolean functions including the algebraic degree and the nonlinearity. Very often it is rather easy to find Boolean functions that are optimal with respect to one of these measures, as for example bent functions. However, the best Boolean functions with respect to one measure can be weak with respect to other measures. Hence Boolean functions that guarantee good behavior with respect to all or at least many such measures are in high demand. In this sense, finding a good cryptographic Boolean function is somehow like finding a spouse where also a trade-off between different desirable features is needed. (The wives of the authors are exceptions since they both have all desirable features.) We will explain this more carefully—for Boolean functions and not for spouses where we recommend, say, the books of the world-famous relationship counselor John Gray.

Each Boolean function f of n variables can be uniquely represented by a polynomial

$$P(x_1, \dots, x_n) = \sum_{i_1,\dots,i_n=0}^{1} a_{i_1,\dots,i_n} x_1^{i_1} \cdots x_n^{i_n} \in \mathbb{F}_2[x_1, \dots, x_n], \tag{6.16}$$

that is,

$$f(c_1, \dots, c_n) = P(c_1, \dots, c_n) \quad \text{for all } (c_1, \dots, c_n) \in \mathbb{F}_2^n$$

with the convention $0^0 = 1 \in \mathbb{F}_2$. This representation is called the *algebraic normal form* (*ANF*) of f. The *algebraic degree* of a Boolean function f with ANF (6.16) is defined by

$$\deg(f) = \max\{i_1 + \cdots + i_n : a_{i_1,\dots,i_n} = 1\},$$

where $\deg(f) = 0$ if all $a_{i_1,\dots,i_n} = 0$.

Boolean functions of small algebraic degree are predictable and the coefficients of their ANF can be determined from a small system of linear equations. More precisely, a Boolean function f of n variables and of algebraic degree d has at most $1 + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{d}$ nonzero coefficients and, for example, the values of f at all $(c_1, \dots, c_n) \in \mathbb{F}_2^n$ with at most d coordinates $c_i = 1$ define such a system of linear equations. Hence a large algebraic degree is desirable for a cryptographic Boolean function. Moreover, for $n \geq 2$ the number of Boolean functions of n variables and of algebraic degree at most $n-2$ is $2^{\sum_{i=0}^{n-2}\binom{n}{i}} = 2^{2^n-n-1}$, which is negligible compared to the total number 2^{2^n} of all Boolean functions of n variables if n is large. In this sense, almost all Boolean functions of n variables are of algebraic degree $n-1$ or n.

The *nonlinearity* $\mathcal{NL}(f)$ of a Boolean function f of n variables is defined by

$$\mathcal{NL}(f) = 2^{n-1} - \frac{1}{2} \max_{\mathbf{u} \in \mathbb{F}_2^n} |\hat{F}(\mathbf{u})|.$$

It is a measure for how different f is from all affine Boolean functions (compare with Remark 6.4.23), and thus a Boolean function of small nonlinearity is again predictable. We infer from Corollary 6.4.25 that

$$\mathcal{NL}(f) \leq 2^{n-1} - 2^{n/2-1}$$

for all Boolean functions f of n variables, and this upper bound is attained exactly for bent functions. However, the algebraic degree of a bent function is at most $n/2$ for $n \geq 4$.

Now we present a number-theoretic construction of a Boolean function of n variables with algebraic degree at least $n-1$ for which we can still prove an interesting lower bound on its nonlinearity.

Let p be an odd prime number and put $n = \lfloor \log_2 p \rfloor$. We introduce the Boolean function g of n variables given by

$$g(c_1, \dots, c_n) = \begin{cases} 0 & \text{if } c \text{ is a quadratic residue modulo } p, \\ 1 & \text{otherwise,} \end{cases} \tag{6.17}$$

where $c = c_1 + 2c_2 + \cdots + 2^{n-1}c_n$ with $c_1, \ldots, c_n \in \{0, 1\}$. We note that 2 is a quadratic residue modulo p if and only if $p \equiv \pm 1 \pmod 8$ (see [152, Theorem 3.3]).

Theorem 6.4.29 *Let p be a prime number with $p \equiv \pm 3 \pmod 8$ and let g be the Boolean function defined by* (6.17). *Then* $\deg(g) \geq n - 1$.

Proof. The case $n = 1$ is trivial, and so we can assume that $n \geq 2$. Put

$$h(c_1, \ldots, c_{n-1}) := g(0, c_1, \ldots, c_{n-1}) + g(c_1, \ldots, c_{n-1}, 0)$$

for all $(c_1, \ldots, c_{n-1}) \in \mathbb{F}_2^{n-1}$. If c is a quadratic residue modulo p, then $2c$ is a quadratic nonresidue modulo p and vice versa since by assumption 2 is a quadratic nonresidue modulo p. Hence one of the two summands in the definition of h is 1 and the other one is 0. Therefore

$$h(c_1, \ldots, c_{n-1}) = \begin{cases} 1 & \text{if } (c_1, \ldots, c_{n-1}) \neq (0, \ldots, 0), \\ 0 & \text{if } (c_1, \ldots, c_{n-1}) = (0, \ldots, 0), \end{cases}$$

which can be written in the form

$$h(c_1, \ldots, c_{n-1}) = \prod_{i=1}^{n-1} (1 + c_i) + 1.$$

It follows that $\deg(g) \geq \deg(h) = n - 1$. □

Theorem 6.4.30 *Let p be an odd prime number and let g be the Boolean function defined by* (6.17). *Then*

$$\mathcal{NL}(g) > 2^{n-1} - (n+2)^{1/2} 2^{7n/8} - \frac{1}{2}.$$

Proof. The bound is trivial for $n = 1$ and $n = 2$, and so we can assume that $n \geq 3$. Let η be the quadratic character of $\mathbb{F}_p$ (see Remark 1.4.53). Since

$$\eta(v) = (-1)^{g(v_1, \ldots, v_n)} \qquad \text{for } 1 \leq v \leq 2^n - 1,$$

the Hadamard transform $\hat{G}$ of g can be written in the form

$$\hat{G}(\mathbf{u}) = \sum_{\mathbf{v} \in \mathbb{F}_2^n} (-1)^{\mathbf{u} \cdot \mathbf{v} + g(\mathbf{v})} = \sum_{v=0}^{2^n - 1} \eta(v) (-1)^{\mathbf{u} \cdot \mathbf{v}} + (-1)^{g(\mathbf{0})},$$

where $\mathbf{u} \in \mathbb{F}_2^n$ and $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{F}_2^n$ if $v = v_1 + 2v_2 + \cdots + 2^{n-1}v_n$. We introduce the sum

$$S(\mathbf{u}) = \sum_{v=0}^{2^n - 1} \eta(v) (-1)^{\mathbf{u} \cdot \mathbf{v}}$$

and note that $|\hat{G}(\mathbf{u})| \leq |S(\mathbf{u})| + 1$. We put

$$k = \left\lceil \frac{3}{4}(n+1) \right\rceil \leq n, \quad N = 2^k, \quad M = 2^{n-k}.$$

With $\mathbf{w} = (u_1, \ldots, u_k)$ and $\mathbf{x} = (u_{k+1}, \ldots, u_n)$, that is, $\mathbf{u} = (\mathbf{w}, \mathbf{x}) \in \mathbb{F}_2^k \times \mathbb{F}_2^{n-k}$, we can write

$$S(\mathbf{u}) = \sum_{a=0}^{N-1} \sum_{b=0}^{M-1} \eta(a + Nb)(-1)^{\mathbf{a}\cdot\mathbf{w}+\mathbf{b}\cdot\mathbf{x}},$$

where $\mathbf{a} \in \mathbb{F}_2^k$ corresponds to the integer a and $\mathbf{b} \in \mathbb{F}_2^{n-k}$ corresponds to the integer b in the way we have seen before. Then

$$|S(\mathbf{u})| \leq \sum_{a=0}^{N-1} \Big| \sum_{b=0}^{M-1} \eta(a + Nb)(-1)^{\mathbf{b}\cdot\mathbf{x}} \Big|.$$

An application of the Cauchy-Schwarz inequality yields

$$\begin{aligned} |S(\mathbf{u})|^2 &\leq \Big(\sum_{a=0}^{N-1} 1 \cdot \Big| \sum_{b=0}^{M-1} \eta(a+Nb)(-1)^{\mathbf{b}\cdot\mathbf{x}} \Big| \Big)^2 \\ &\leq N \sum_{a=0}^{N-1} \Big| \sum_{b=0}^{M-1} \eta(a+Nb)(-1)^{\mathbf{b}\cdot\mathbf{x}} \Big|^2 \\ &\leq N \sum_{b_1,b_2=0}^{M-1} \Big| \sum_{a=0}^{N-1} \eta((a+Nb_1)(a+Nb_2)) \Big|. \end{aligned}$$

For the M ordered pairs (b_1, b_2) with $b_1 = b_2$, the absolute value of the inner sum is trivially bounded by N. For the remaining $M(M-1)$ ordered pairs (b_1, b_2) with $b_1 \neq b_2$, the inner sum is of the form of the character sum in (6.8), but for parts of the period. By a standard bound (see [182, Lemma 3.4]), the absolute value of the inner sum is at most $2p^{1/2} \log p$. Collecting everything gives

$$\begin{aligned} |S(\mathbf{u})|^2 &\leq N(MN + 2M^2 p^{1/2} \log p) \leq 2^k(2^n + 2^{2n-2k+1} 2^{(n+1)/2}(n+1)) \\ &\leq 2^{n+k}(n+2) < 4 \cdot 2^{7n/4}(n+2). \end{aligned}$$

The proof is completed by recalling that $|\hat{G}(\mathbf{u})| \leq |S(\mathbf{u})| + 1$ for all $\mathbf{u} \in \mathbb{F}_2^n$. □

There are several other cryptographic measures for Boolean functions including the nonlinearity of higher order, the m-resiliency, and the algebraic thickness. For more details on Boolean functions and their cryptographic measures, we refer to the monograph [33] and the survey article [18].

6.5 Number theory and quantum computation

6.5.1 The hidden subgroup problem

A quantum computer performs operations on data based on quantum-mechanical phenomena. Although small-scale experiments for quantum computation have already been carried out, such as factoring the number 15, large-scale quantum computers are currently out of reach. However, several efficient algorithms for large-scale quantum computers have already been developed, including the Shor algorithm for factoring integers. In this subsection, we take a quantum algorithm for solving the hidden subgroup problem as a black-box, that is, the specifics of the quantum algorithm are not our concern, and we explain how it can be used to resolve the factoring problem and the discrete logarithm problem. A detailed introduction to quantum computation and quantum information theory is out of the scope of this book and we refer instead to the monograph [150].

The *hidden subgroup problem* can be phrased as follows: let f be a function from an abelian group G to a finite set X such that f is constant on the cosets of a subgroup K of G and has distinct values on different cosets; then find K from evaluations of f. We think of f as hiding the subgroup K. Although there is no classical algorithm known for solving this problem efficiently, a quantum computer could crack the hidden subgroup problem. We emphasize again that we use such an algorithm only as a black-box and we refer to [150, Section 5.4] for the details.

The quantum computation part of the celebrated algorithm of Shor [180] for factoring integers solves the following *period-finding problem* efficiently on a quantum computer: let f be any periodic function from $\mathbb{Z}$ into a finite set X without repetition in a period; then find $r \in \mathbb{N}$ with $f(m+r) = f(m)$ for all $m \in \mathbb{Z}$. Here the hidden subgroup of $\mathbb{Z}$ is $K = \{nr : n \in \mathbb{Z}\}$. If X is a finite abelian group with the multiplicative notation, $a \in X$ an element of order r, and $f(m) = a^m$, then we get an *order-finding problem*.

Both the factoring problem and the discrete logarithm problem can be reduced to an instance of the hidden subgroup problem. We discuss the factoring problem first and we consider only the situation that arises when one attempts to break the RSA public-key cryptosystem (see Subsection 2.3.2).

Algorithm 6.5.1 (Shor Algorithm) Let $n = pq$ be the product of two different odd prime numbers p and q. Find p and q.
Step 1: choose a random integer a with $1 \leq a < n$, where we may assume that $\gcd(a, n) = 1$ since otherwise $\gcd(a, n)$ is either p or q.
Step 2: use the (quantum) order-finding algorithm to compute the multiplicative order r of a modulo n.
Step 3: if r is even and $a^{r/2} \not\equiv -1 \pmod{n}$, then $\gcd(a^{r/2} - 1, n)$ and $\gcd(a^{r/2} + 1, n)$ are the prime factors of n; otherwise return to Step 1.

The confirmation of the efficiency of this probabilistic algorithm is based on the following theorem. We recall from Example 1.3.7 that R_m denotes the finite abelian group formed by the integers a with $0 \le a < m$ and $\gcd(a, m) = 1$ under multiplication modulo m, where m is any positive integer.

Theorem 6.5.2 *Let $n = pq$ be the product of two different odd prime numbers $p = 2^{s_1}d_1 + 1$ and $q = 2^{s_2}d_2 + 1$ with integers $1 \le s_1 \le s_2$ and odd integers d_1 and d_2. Suppose that $a \in R_n$ is chosen uniformly at random. Then the probability that the multiplicative order r of a modulo n is even and $a^{r/2} \not\equiv -1 \pmod{n}$ is*

$$1 - \frac{4^{s_1} + 2}{3 \cdot 2^{s_1+s_2}} \ge \frac{1}{2}.$$

Proof. First we note that for every positive divisor t of $p - 1$, the congruence $a^t \equiv 1 \pmod{p}$ has exactly t solutions $a \in R_p$, namely $a \equiv g^{j(p-1)/t} \pmod{p}$, $j = 0, 1, \ldots, t - 1$, if g is a primitive root modulo p. Moreover, if t is a positive divisor of $(p - 1)/2$, then $a^t \equiv -1 \pmod{p}$ has exactly t solutions $a \in R_p$, namely $a \equiv g^{j(p-1)/t+(p-1)/2t} \pmod{p}$, $j = 0, 1, \ldots, t - 1$.

We show that the probability that either r is odd or r is even and $a^{r/2} \equiv -1 \pmod{n}$ is

$$\frac{1}{2^{s_1+s_2}} + \frac{4^{s_1} - 1}{3 \cdot 2^{s_1+s_2}} = \frac{4^{s_1} + 2}{3 \cdot 2^{s_1+s_2}} \le \frac{1}{3} + \frac{2}{3 \cdot 4^{s_1}} \le \frac{1}{2}.$$

Let r_1 and r_2 be the multiplicative orders of a modulo p and modulo q, respectively. First we prove that the probability that r is odd is $2^{-s_1-s_2}$. The elements of odd order in R_p are exactly the d_1 elements $a \in R_p$ with $a^{d_1} \equiv 1 \pmod{p}$, and so the probability that r_1 is odd is $d_1/(p - 1) = 2^{-s_1}$. Similarly, the probability that r_2 is odd is $d_2/(q - 1) = 2^{-s_2}$. Now r is the least common multiple of r_1 and r_2. Hence r is odd if and only if both r_1 and r_2 are odd, which occurs with probability $2^{-s_1-s_2}$.

Now we show that the probability that r is even and $a^{r/2} \equiv -1 \pmod{n}$ is

$$\frac{4^{s_1} - 1}{3 \cdot 2^{s_1+s_2}}.$$

If $r = 2^s d$ is even with an integer $s \ge 1$ and an odd integer d, and if $a^{r/2} \equiv -1 \pmod{n}$, then $a^{r/2} \equiv -1 \pmod{p}$. Hence r_1 does not divide $r/2$, but r_1 divides r and 2^s is the largest power of 2 dividing r_1. Note that $1 \le s \le s_1$ since r_1 divides $p - 1$. The elements $a \in R_p$ of such an order are characterized by $a^{2^{s-1}d_1} \equiv -1 \pmod{p}$. Their number is $2^{s-1}d_1$ and their probability is 2^{s-s_1-1}. Similarly, 2^s is also the largest power of 2 dividing r_2 and the elements $a \in R_q$ of such an order are characterized by $a^{2^{s-1}d_2} \equiv -1 \pmod{q}$. Their number is $2^{s-1}d_2$ and their probability is 2^{s-s_2-1}. So the probability that $a \in R_n$ has an order of the form $r = 2^s d$ is

$$\frac{4^{s-1}}{2^{s_1+s_2}}, \qquad s = 1, \ldots, s_1,$$

which sums up to

$$\frac{4^{s_1} - 1}{3 \cdot 2^{s_1+s_2}},$$

and so we are done. □

Remark 6.5.3 You have certainly figured out the conclusion in Step 3 of Algorithm 6.5.1, but since this point is important, we flog a dead horse and write down the argument. If r is even, then since $a^r \equiv 1 \pmod{n}$, we know that $a^r - 1 = (a^{r/2} - 1)(a^{r/2}+1)$ is divisible by n. If $a^{r/2} \not\equiv -1 \pmod{n}$, then since also $a^{r/2} \not\equiv 1 \pmod{n}$, one of the prime factors of n must divide $a^{r/2} - 1$ and the other one $a^{r/2} + 1$.

Example 6.5.4 For $n = 21$ there are three elements $a \in R_{21}$ of odd order r and three elements $a \in R_{21}$ with even order r and $a^{r/2} \equiv -1 \pmod{21}$. Since $|R_{21}| = \phi(21) = 12$, the probability to get a random element $a \in R_{21}$ with even order r and $a^{r/2} \not\equiv -1 \pmod{21}$ is exactly $\frac{1}{2}$. The boldface numbers in the following table are warning signs that highlight cases where one of the conditions in Step 3 of Algorithm 6.5.1 is not satisfied.

a	1	2	4	5	8	10	11	13	16	17	19	20
r	**1**	6	**3**	6	2	6	6	2	**3**	6	6	2
$a^{r/2} \pmod{21}$	–	8	–	$-\mathbf{1}$	8	-8	8	-8	–	$-\mathbf{1}$	-8	$-\mathbf{1}$
$\gcd(a^{r/2}+1, 21)$	–	3	–	21	3	7	3	7	–	21	7	21
$\gcd(a^{r/2}-1, 21)$	–	7	–	1	7	3	7	3	–	1	3	1

Another instance of the hidden subgroup problem solves the discrete logarithm problem. The basic idea of this algorithm also goes back to the paper [180].

Algorithm 6.5.5 Let q be a prime power, let $a \in \mathbb{F}_q^*$ be an element of order r, and suppose we know that a given $b \in \mathbb{F}_q^*$ satisfies $b = a^s$ with $s \in \mathbb{Z}$ and $0 \leq s < r$. Find s.
Step 1: take $f : Z_r^2 \to \mathbb{F}_q$ with $f(h_1, h_2) = b^{h_1} a^{h_2}$ for all $(h_1, h_2) \in Z_r^2$ and find the (hidden) subgroup

$$K = \{(k, -ks) : k \in Z_r\}$$

of Z_r^2 using the (quantum) algorithm for solving the hidden subgroup problem.
Step 2: choose an arbitrary element $(k_1, k_2) \in K$ with $\gcd(k_1, r) = 1$ and determine s from the congruence $k_1 s \equiv -k_2 \pmod{r}$.

Remark 6.5.6 Since $f(h_1+k, h_2-ks) = b^{h_1+k}a^{h_2-ks} = b^{h_1}b^k a^{h_2} b^{-k} = f(h_1, h_2)$ for all $k \in Z_r$, the function f is constant on each coset $(h_1, h_2) + K$ of K. It remains to show that two ordered pairs $(h_1, h_2), (j_1, j_2) \in Z_r^2$ with $f(h_1, h_2) = f(j_1, j_2)$ belong to the same coset of K. From $b^{h_1}a^{h_2} = b^{j_1}a^{j_2}$ we get, with $b = a^s$ and since r is the order of a, that

$$h_1 s + h_2 \equiv j_1 s + j_2 \pmod{r}.$$

Hence in the group Z_r^2 this yields the identity

$$(h_1 - j_1, h_2 - j_2) = (h_1 - j_1, -(h_1 - j_1)s) \in K,$$

and thus $(h_1, h_2) + K = (j_1, j_2) + K$.

Example 6.5.7 Take $q = 7$, $a = 3$, $b = 4$, and so $r = 6$. Then $f : Z_6^2 \to \mathbb{F}_7$ is given by $f(h_1, h_2) \equiv 4^{h_1} 3^{h_2} \pmod 7$ for all $(h_1, h_2) \in Z_6^2$. Here $K = \{(0,0), (1,2), (2,4), (3,0), (4,2), (5,4)\}$ consists of the elements $(c, d) \in Z_6^2$ with

$$f(h_1 + c, h_2 + d) \equiv 4^{h_1+c} 3^{h_2+d} \equiv 4^{h_1} 3^{h_2} \equiv f(h_1, h_2) \pmod 7,$$

that is, $4^c 3^d \equiv 1 \pmod 7$. Choose any $(k_1, k_2) \in K$ with $\gcd(k_1, 6) = 1$, say $(k_1, k_2) = (5, 4)$, and determine s from the congruence $5s \equiv -4 \pmod 6$. This yields $s = 4$. You can check that $a^s \equiv 3^4 \equiv 4 \equiv b \pmod 7$.

The upshot of Algorithms 6.5.1 and 6.5.5 is of course that as soon as large-scale quantum computers are available, then cryptographic schemes based on the presumed difficulty of factoring integers or on the presumed difficulty of solving the discrete logarithm problem will be seriously compromised. But cryptographers are visionary people and they have thought for many years about alternative schemes that may stand tall against the onslaught of quantum computers. An entire branch of cryptography called post-quantum cryptography is devoted to the design of such alternative schemes (see the book [11]). Examples of cryptographic schemes that will, as far as one can tell at present, survive quantum computers are lattice-based cryptosystems and code-based cryptosystems; the latter were briefly discussed in Section 3.6.

6.5.2 Mutually unbiased bases

Mutually unbiased bases were introduced in the literature on quantum mechanics by Schwinger [178]. They are important not only for quantum physics, but also for applications to quantum information theory. Mutually unbiased bases are collections of orthonormal bases of a complex vector space with a characteristic property described in Definition 6.5.8 below.

The setting is the n-dimensional complex vector space $\mathbb{C}^n$ with $n \geq 2$. This vector space is endowed with the *Hermitian inner product*

$$\langle \mathbf{y} | \mathbf{z} \rangle = \sum_{j=1}^{n} \overline{y_j} z_j$$

for all $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{C}^n$ and $\mathbf{z} = (z_1, \ldots, z_n) \in \mathbb{C}^n$, where the bar denotes complex conjugation. By the way, this is the definition of the Hermitian inner product that is used in the theory of mutually unbiased bases and stems from quantum mechanics. The standard definition in the mathematical literature takes the complex conjugate thereof. A basis $B = \{\mathbf{w}_1, \dots, \mathbf{w}_n\}$ of $\mathbb{C}^n$ is an *orthonormal basis* of $\mathbb{C}^n$ if for $1 \le j, k \le n$,

$$\langle \mathbf{w}_j | \mathbf{w}_k \rangle = \begin{cases} 1 & \text{if } j = k, \\ 0 & \text{if } j \ne k. \end{cases}$$

For instance, the standard basis $S = \{\mathbf{s}_1, \dots, \mathbf{s}_n\}$ of $\mathbb{C}^n$ is an orthonormal basis, where $\mathbf{s}_j$, $j = 1, \dots, n$, has jth coordinate equal to 1 and all other coordinates equal to 0.

Definition 6.5.8 Two orthonormal bases B and B' of $\mathbb{C}^n$ are *mutually unbiased* if

$$|\langle \mathbf{w} | \mathbf{w}' \rangle| = \frac{1}{\sqrt{n}} \quad \text{for all } \mathbf{w} \in B \text{ and } \mathbf{w}' \in B'.$$

For an integer $m \ge 1$, a collection $B_0, B_1, \dots, B_m$ of $m+1$ orthonormal bases of $\mathbb{C}^n$ is *mutually unbiased* if B_h and B_i are mutually unbiased for $0 \le h < i \le m$.

It is known that any collection of mutually unbiased bases of $\mathbb{C}^n$ can contain at most $n+1$ orthonormal bases of $\mathbb{C}^n$. Maximal collections of $n+1$ mutually unbiased bases of $\mathbb{C}^n$ are of considerable interest.

Example 6.5.9 For $n = 2$, the orthonormal bases B_0, B_1, B_2 of $\mathbb{C}^2$ given by

$$\begin{aligned} B_0 &= \{(1,0), (0,1)\}, \\ B_1 &= \left\{ \frac{1}{\sqrt{2}}(1,1), \frac{1}{\sqrt{2}}(1,-1) \right\}, \\ B_2 &= \left\{ \frac{1}{\sqrt{2}}(1,\mathrm{i}), \frac{1}{\sqrt{2}}(1,-\mathrm{i}) \right\} \end{aligned}$$

form a maximal collection of three mutually unbiased bases of $\mathbb{C}^2$. Here $\mathrm{i} = \sqrt{-1}$ is as usual the imaginary unit.

Believe it or not, maximal collections of mutually unbiased bases can be constructed with the help of finite fields, although the fields $\mathbb{C}$ and $\mathbb{F}_q$ are of a quite different nature. This is another demonstration of the unity of mathematics.

Theorem 6.5.10 *Let $n = q$ be a power of an odd prime and let χ be a nontrivial additive character of $\mathbb{F}_q$. For every $h \in \mathbb{F}_q$, define $B_h = \{\mathbf{w}_{h,k}\}_{k \in \mathbb{F}_q}$ by*

$$\mathbf{w}_{h,k} = \frac{1}{\sqrt{q}} \left(\chi(ha^2 + ka) \right)_{a \in \mathbb{F}_q} \in \mathbb{C}^q \qquad \text{for all } k \in \mathbb{F}_q.$$

Then the standard basis S of $\mathbb{C}^q$ and the B_h for $h \in \mathbb{F}_q$ form a maximal collection of $q+1$ mutually unbiased bases of $\mathbb{C}^q$.

Proof. We already know that the standard basis S of $\mathbb{C}^q$ is an orthonormal basis of $\mathbb{C}^q$. Next we show that B_h is an orthonormal basis of $\mathbb{C}^q$ for every $h \in \mathbb{F}_q$. This follows from

$$\langle \mathbf{w}_{h,j} | \mathbf{w}_{h,k} \rangle = \frac{1}{q} \sum_{a \in \mathbb{F}_q} \overline{\chi(ha^2 + ja)} \chi(ha^2 + ka) = \frac{1}{q} \sum_{a \in \mathbb{F}_q} \chi((k-j)a)$$

for all $j, k \in \mathbb{F}_q$ and the orthogonality relation (1.9). It is trivial that S and each B_h with $h \in \mathbb{F}_q$ are mutually unbiased. Finally, we consider B_h and B_i with $h, i \in \mathbb{F}_q$ and $h \neq i$. Then

$$|\langle \mathbf{w}_{h,j} | \mathbf{w}_{i,k} \rangle| = \frac{1}{q} \Big| \sum_{a \in \mathbb{F}_q} \chi((i-h)a^2 + (k-j)a) \Big| = \frac{1}{q} \Big| \sum_{b \in \mathbb{F}_q} \chi((i-h)b^2) \Big|$$

for all $j, k \in \mathbb{F}_q$. In the last step we used the usual trick of completing the square (which works since q is odd) and a suitable change of the summation variable. For the absolute value of the last sum we get with $c = i - h \in \mathbb{F}_q^*$,

$$\Big| \sum_{b \in \mathbb{F}_q} \chi(cb^2) \Big|^2 = \sum_{a,b \in \mathbb{F}_q} \chi(c(a^2 - b^2)) = \sum_{b \in \mathbb{F}_q} \sum_{a \in \mathbb{F}_q} \chi(c((a+b)^2 - b^2))$$

$$= \sum_{b \in \mathbb{F}_q} \sum_{a \in \mathbb{F}_q} \chi(c(a^2 + 2ab)) = \sum_{a \in \mathbb{F}_q} \chi(ca^2) \sum_{b \in \mathbb{F}_q} \chi(2cab) = q,$$

and the result follows. Note that in the last step we again used (1.9) and the fact that q is odd. □

Further constructions of maximal collections of mutually unbiased bases using finite fields can be found in [80]. There is also a construction in [80] for dimensions n that are powers of 2 based on so-called Galois rings, which are algebraic structures that are somewhat more general than finite fields. Thus, maximal collections of $n+1$ mutually unbiased bases of $\mathbb{C}^n$ are known for all prime powers n. It is conjectured that there are dimensions n for which collections of $n+1$ mutually unbiased bases of $\mathbb{C}^n$ do not exist. There is particularly strong evidence for this in the case $n = 6$ (see again [80]).

6.6 Primality tests

6.6.1 Fermat test and Carmichael numbers

Large prime numbers are needed in several applications including cryptography (see Chapter 2) and pseudorandom number generation (see Chapter 5). This raises the issue of how to decide whether a given large integer is a prime number or a composite number, and this is what primality tests are all about. Because of their importance for areas such as cryptography, there is an extensive literature on primality tests. Exemplary treatments of the subject are given in the standard monographs of Bach and Shallit [6], Crandall and Pomerance [30], and Riesel [169]. Just to satisfy your curiosity, we briefly discuss some classical primality tests and we mention an important more recent achievement in Remark 6.6.12.

There is a basic dichotomy between probabilistic primality tests (also called pseudoprime tests) and deterministic primality tests. In a probabilistic primality test, as in any probabilistic algorithm, we are allowed to make random choices in various steps of the algorithm. A typical probabilistic algorithm (though for factoring and not for primality testing) is Algorithm 6.5.1. Probabilistic primality tests tend to be faster, but there is no absolute guarantee of success. Indeed, there can be composite numbers that pass the primality test even after many random choices in the algorithm.

A simple deterministic primality test is based on the observation that is as old as the hills, namely that a composite number n has a divisor d with $2 \leq d \leq n^{1/2}$. Thus, we test whether any of the integers $2, 3, \ldots, \lfloor n^{1/2} \rfloor$ divides the given integer $n \geq 4$, and if this is not the case, then we know for sure that n is a prime number. However, this primality test is futile for integers n of cryptographic relevance, that is, for n of the order of magnitude 10^{150} or even larger. We need vastly more efficient algorithms for such n.

When designing a primality test, whether deterministic or probabilistic, it is a good idea to look for a condition that a prime number must necessarily satisfy. If this condition fails to hold for an integer $n \geq 2$, then we can infer that n is composite. A simple and elegant necessary condition for primality is given by Fermat's little theorem (see Corollary 1.2.16), and this is the basis for the following primality test. Recall that Fermat's little theorem says that if n is a prime number, then $a^{n-1} \equiv 1 \pmod{n}$ for all $a \in \mathbb{Z}$ with $\gcd(a, n) = 1$. Thus, if an arbitrary integer $n \geq 2$ is given and we can find an integer a with $a^{n-1} \not\equiv 1 \pmod{n}$ and $\gcd(a, n) = 1$, then n must be composite. The *Fermat test* proceeds by randomly picking integers a with $\gcd(a, n) = 1$, where we can assume also that $1 \leq a \leq n - 1$, and checking whether $a^{n-1} \equiv 1 \pmod{n}$ or not. If n is very large, then the power a^{n-1} can be computed by the efficient square-and-multiply algorithm (see Algorithm 2.3.9). We may of course be tempted to conclude that if $a^{n-1} \equiv 1 \pmod{n}$ for all $a \in \mathbb{Z}$ with $\gcd(a, n) = 1$, then n is a prime number. But we are in for a bad surprise!

Example 6.6.1 Let $n = 561 = 3 \cdot 11 \cdot 17$. Now we take an arbitrary $a \in \mathbb{Z}$ with $\gcd(a, 561) = 1$. Then $\gcd(a, 3) = \gcd(a, 11) = \gcd(a, 17) = 1$, and so Fermat's little theorem yields $a^2 \equiv 1 \pmod{3}$, $a^{10} \equiv 1 \pmod{11}$, and $a^{16} \equiv 1 \pmod{17}$. By raising these congruences to suitable powers, we obtain $a^{560} \equiv 1 \pmod{3}$, $a^{560} \equiv 1 \pmod{11}$, and $a^{560} \equiv 1 \pmod{17}$, and so $a^{560} \equiv 1 \pmod{561}$ for all $a \in \mathbb{Z}$ with $\gcd(a, 561) = 1$. This means that no matter which integer a with $\gcd(a, n) = 1$ we try, the condition

$a^{n-1} \equiv 1 \pmod{n}$ is satisfied for $n = 561$, but nevertheless 561 is a composite number. Thus, the Fermat test fails for this n. We give a special name to these bad guys n.

Definition 6.6.2 A composite number n satisfying $a^{n-1} \equiv 1 \pmod{n}$ for all $a \in \mathbb{Z}$ with $\gcd(a, n) = 1$ is called a *Carmichael number*.

Therefore 561 is a Carmichael number and it is in fact the smallest Carmichael number. Here are some properties of Carmichael numbers.

Proposition 6.6.3 *Every Carmichael number is squarefree, that is, it is not divisible by the square of a prime number.*

Proof. Assume that n is a Carmichael number with p^2 dividing n for some prime number p. The multiplicative group R_{p^2} is cyclic by [152, Theorem 2.41], and so there exists an element $g \in R_{p^2}$ of order $|R_{p^2}| = \phi(p^2) = p(p-1)$. By the Chinese remainder theorem (see Theorem 1.2.9), we can choose $a \in \mathbb{Z}$ with $a \equiv g \pmod{p^2}$ and $\gcd(a, n) = 1$. Then $1 \equiv a^{n-1} \equiv g^{n-1} \pmod{p^2}$, and so $p(p-1)$ divides $n-1$ by Lemma 1.3.10. In particular, p divides $n-1$, a contradiction. □

Theorem 6.6.4 *Let n be a squarefree composite number. Then n is a Carmichael number if and only if $p-1$ divides $n-1$ for every prime factor p of n.*

Proof. If $p-1$ divides $n-1$ for every prime factor p of n, then n is a Carmichael number by the same argument as in Example 6.6.1. Conversely, suppose that n is a Carmichael number and let p be a prime factor of n. For a primitive root g modulo p, we choose $a \in \mathbb{Z}$ with $a \equiv g \pmod{p}$ and $a \equiv 1 \pmod{n/p}$ by the Chinese remainder theorem. Then $\gcd(a, n) = 1$, hence $g^{n-1} \equiv a^{n-1} \equiv 1 \pmod{p}$, and so $p-1$ divides $n-1$ by Lemma 1.3.10. □

By using Theorem 6.6.4, you can check again that $561 = 3 \cdot 11 \cdot 17$ is a Carmichael number. In the same way, you can verify that $1729 = 7 \cdot 13 \cdot 19$ is a Carmichael number. Fans of mathematical anecdotes know 1729 as Hardy's cab number. Here is the story. The famous number theorist Hardy, the first author of the book [62] and the author of *A Mathematician's Apology* (see the preface of the present book), took a cab to visit his hospitalized friend Ramanujan, another famous number theorist. Being an avid collector of numbers, Hardy noted the cab number and told Ramanujan by way of conversation that he had come in a cab with the dull number 1729. Ramanujan, an advocate of equal opportunity for all numbers, protested and pointed out that 1729 is interesting because it is the smallest positive integer that can be expressed in two different ways as the sum of two cubes of positive integers, namely $1729 = 12^3 + 1^3 = 10^3 + 9^3$. But Ramanujan failed to mention that 1729 is interesting as well because it is the third Carmichael number (in the natural order). His excuse is that he was sick. By the way, the second Carmichael number (in the natural order) is $1105 = 5 \cdot 13 \cdot 17$.

You have noticed that each of the three concrete Carmichael numbers we have mentioned has three distinct prime factors. There is a general result behind this observation.

Proposition 6.6.5 *Every Carmichael number has at least three distinct prime factors.*

Proof. Each Carmichael number n is squarefree by Proposition 6.6.3. Assume that $n = pq$ with two primes $p < q$. Then $q \equiv 1 \pmod{q-1}$, hence $n - 1 \equiv p - 1 \not\equiv 0 \pmod{q-1}$, and so $q-1$ does not divide $n-1$. This is a contradiction to Theorem 6.6.4. □

One could have the vague hope that the existence of Carmichael numbers is a phenomenon for small integers and that sufficiently large composite numbers are not Carmichael numbers. But it was shown in the deep paper of Alford, Granville, and Pomerance [2] that there are infinitely many Carmichael numbers. Thus, we have to live with Carmichael numbers when performing the Fermat test. If we know that a given composite number is not a Carmichael number, then the analysis of the Fermat test is easy.

Proposition 6.6.6 *If the composite number n is not a Carmichael number, then there are at most $\phi(n)/2$ different integers a with $1 \leq a \leq n-1$, $\gcd(a, n) = 1$, and $a^{n-1} \equiv 1 \pmod{n}$.*

Proof. The set

$$T_n = \{a \in R_n : a^{n-1} \equiv 1 \pmod{n}\} \subseteq R_n = \{a \in Z_n : \gcd(a, n) = 1\}$$

is a subgroup of the multiplicative group R_n. Since n is not a Carmichael number, T_n is a proper subgroup of R_n, and the result follows from Lagrange's theorem (see Theorem 1.3.21). □

Remark 6.6.7 If the composite number n is not a Carmichael number, then for a random choice of $a \in \mathbb{Z}$ with $\gcd(a, n) = 1$, the probability that $a^{n-1} \not\equiv 1 \pmod{n}$ is at least $\frac{1}{2}$ according to Proposition 6.6.6. Thus, after a few random choices of a it is detected with high probability that n is composite. For a very large n which is not a Carmichael number, we can use only a small fraction of the candidates $a \in R_n$ in practice, and even if for all these a we have $a^{n-1} \equiv 1 \pmod{n}$, there is no guarantee that n is a prime number, but there is a high probability for it. In this case, the experts speak of a "probable prime".

6.6.2 Solovay-Strassen test

Since the Fermat test fails for the infinitely many Carmichael numbers, we need a more sophisticated primality test. The test suggested by Solovay and Strassen [191] is based on the theory of quadratic residues. Since every even integer greater than 3 is composite, we can assume that the number n to be tested for primality is odd.

The necessary condition for primality that we use now is obtained from Proposition 1.2.23: if n is an odd prime number, then $a^{(n-1)/2} \equiv \left(\frac{a}{n}\right) \pmod{n}$ for all $a \in \mathbb{Z}$ with $\gcd(a, n) = 1$. The *Solovay-Strassen test* for an odd integer $n \geq 3$ is performed by randomly picking integers a with $\gcd(a, n) = 1$ and $1 \leq a \leq n - 1$ and by checking whether $a^{(n-1)/2} \equiv \left(\frac{a}{n}\right) \pmod{n}$ or not. If you are an attentive reader, then you notice that we have not yet defined the symbol $\left(\frac{a}{n}\right)$ for composite numbers n. However, it is a simple step from the Legendre symbol $\left(\frac{a}{p}\right)$ to the *Jacobi symbol* $\left(\frac{a}{n}\right)$ for odd composite numbers $n \geq 3$. We take the canonical factorization $n = \prod_{j=1}^{k} p_j^{e_j}$ of n and then the Jacobi symbol is given by

$$\left(\frac{a}{n}\right) = \prod_{j=1}^{k} \left(\frac{a}{p_j}\right)^{e_j} \qquad \text{for all } a \in \mathbb{Z}.$$

The actual computation of the Jacobi symbol $\left(\frac{a}{n}\right)$ does not use this definition, because we would run around in a circle if for a primality test for n we required the canonical factorization of n. In fact, for large n the Jacobi symbol $\left(\frac{a}{n}\right)$ is efficiently computed by means of the law of quadratic reciprocity without recourse to the canonical factorization of n (see [30, Subsection 2.3.1] and [152, Section 3.3]). The power $a^{(n-1)/2}$ in the Solovay-Strassen test is computed by the square-and-multiply algorithm (see Algorithm 2.3.9).

The big advantage of the Solovay-Strassen test over the Fermat test is that there is no analog of Carmichael numbers for the Solovay-Strassen test. In other words, there is a criterion for primality based on the Solovay-Strassen test, and this criterion can in fact be proved quite easily.

Theorem 6.6.8 *The odd integer $n \geq 3$ is a prime number if and only if $a^{(n-1)/2} \equiv \left(\frac{a}{n}\right) \pmod{n}$ for all $a \in \mathbb{Z}$ with $\gcd(a, n) = 1$.*

Proof. The necessity of the condition follows from Proposition 1.2.23. Conversely, let n be composite with $a^{(n-1)/2} \equiv \left(\frac{a}{n}\right) \pmod{n}$ for all $a \in \mathbb{Z}$ with $\gcd(a, n) = 1$. Then $a^{n-1} \equiv 1 \pmod{n}$ for all such a, and so n is a Carmichael number. Therefore n is squarefree by Proposition 6.6.3. Hence we can write $n = pr$ with an odd prime number p, an odd integer $r \geq 3$, and $\gcd(p, r) = 1$. Let h be a quadratic nonresidue modulo p and choose $a \in \mathbb{Z}$ by the Chinese remainder theorem such that $a \equiv h \pmod{p}$ and $a \equiv 1 \pmod{r}$. Then by assumption,

$$a^{(n-1)/2} \equiv \left(\frac{a}{n}\right) \equiv \left(\frac{a}{p}\right)\left(\frac{a}{r}\right) \equiv \left(\frac{h}{p}\right)\left(\frac{1}{r}\right) \equiv -1 \pmod{n},$$

and so $a^{(n-1)/2} \equiv -1 \pmod{r}$. This is a contradiction to $a \equiv 1 \pmod{r}$. □

Thus, if n is an odd composite number, then the Solovay-Strassen test will ultimately detect this fact. The probabilistic analysis of the Solovay-Strassen test proceeds as in Remark 6.6.7 for the Fermat test, on the basis of the following result. In view of Theorem 6.6.8, the condition that n is not a Carmichael number can be dropped in the present case.

Proposition 6.6.9 *If n is an odd composite number, then there are at most $\phi(n)/2$ different integers a with $1 \le a \le n-1$, $\gcd(a,n) = 1$, and $a^{(n-1)/2} \equiv \left(\frac{a}{n}\right) \pmod{n}$.*

Proof. Use Theorem 6.6.8 and the argument in the proof of Proposition 6.6.6, with T_n replaced by $V_n = \{a \in R_n : a^{(n-1)/2} \equiv \left(\frac{a}{n}\right) \pmod{n}\}$. □

Remark 6.6.10 There is a fascinating connection between the extended Riemann hypothesis (ERH) and the Solovay-Strassen test. By assuming the validity of the ERH, the Solovay-Strassen test can be turned into a deterministic primality test which runs in polynomial time, that is, the number of bit operations required to test an odd integer $n \ge 3$ for primality is at most of the order of magnitude $(\log n)^c$ for some known constant $c > 0$. The crucial number-theoretic result here is the following one: under the ERH, for any odd composite number n there is a positive integer $a \le 2(\log n)^2$ for which either $\gcd(a,n) \ne 1$ or $a^{(n-1)/2} \not\equiv \left(\frac{a}{n}\right) \pmod{n}$. We refer to [6, Section 9.5] and [117] for the details.

A test due to Miller [117] and Rabin [163] refines the Solovay-Strassen test. It is based on the following necessary condition for primality.

Proposition 6.6.11 *Let p be an odd prime number and write $p - 1 = 2^s r$ with an integer $s \ge 1$ and an odd integer r. Then for every $a \in \mathbb{Z}$ with $\gcd(a,p) = 1$, either $a^r \equiv 1 \pmod{p}$ or $a^{2^j r} \equiv -1 \pmod{p}$ for some $j \in \mathbb{Z}$ with $0 \le j \le s-1$.*

Proof. Fermat's little theorem yields $a^{p-1} \equiv a^{2^s r} \equiv 1 \pmod{p}$. If $a^{2^j r} \equiv 1 \pmod{p}$ for some $j \in \mathbb{Z}$ with $1 \le j \le s$, then $a^{2^{j-1} r} \equiv \pm 1 \pmod{p}$. Hence either $a^{2^j r} \equiv -1 \pmod{p}$ for some $0 \le j \le s-1$ or $a^{2^j r} \equiv 1 \pmod{p}$ for all $0 \le j \le s$. □

Actually, the necessary condition in Proposition 6.6.11 is also sufficient for an odd integer $n \ge 3$ to be a prime number (see [6, Lemma 9.4.4]). In other words, we have a criterion for primality analogous to Theorem 6.6.8. The *Miller-Rabin test* for an odd integer $n \ge 3$ proceeds by first writing $n - 1 = 2^s r$ with an integer $s \ge 1$ and an odd integer r, then picking a random integer a with $1 \le a \le n-1$ and $\gcd(a,n) = 1$, and then successively computing $a_0 \equiv a^r \pmod{n}$, $a_1 \equiv a_0^2 \pmod{n}, \ldots, a_k \equiv$

$a_{k-1}^2 \pmod{n}$ until $k = s$ or $a_k \equiv 1 \pmod{n}$. For the probabilistic analysis of the Miller-Rabin test, we need an analog of Proposition 6.6.9. This result for the Miller-Rabin test is actually stronger than for the Solovay-Strassen test, in the sense that the upper bound $\phi(n)/2$ in Proposition 6.6.9 can, in a first step, be changed to the upper bound $(n-1)/4$ for the number of $a \in \mathbb{Z}$ with $1 \leq a \leq n-1$, $\gcd(a,n) = 1$, and either $a^r \equiv 1 \pmod{n}$ or $a^{2^j r} \equiv -1 \pmod{n}$ for some $j \in \mathbb{Z}$ with $0 \leq j \leq s-1$. Unfortunately, the proof of this result is quite involved, and so we refer again to [6, Lemma 9.4.4] for the details. For $n > 9$ the bound $(n-1)/4$ can be improved to $\phi(n)/4$ (see [30, Theorem 3.4.4]). These results mean that, in general, the Miller-Rabin test has a higher chance of detecting composite numbers than the Solovay-Strassen test.

Remark 6.6.12 The *AKS test* named after Agrawal, Kayal, and Saxena [1] is an important breakthrough. It is a deterministic polynomial-time primality test in the sense of Remark 6.6.10, but no unproved hypothesis like the ERH is needed for the complexity analysis of the AKS test. The AKS test is therefore the first unconditional deterministic polynomial-time primality test in history. The starting point of the AKS test is the simple observation that for every prime number p the identity $(x+1)^p = x^p + 1$ holds in the polynomial ring $\mathbb{F}_p[x]$. For every integer $n \geq 2$, we now view Z_n as a finite ring with addition and multiplication modulo n, and we can then form the polynomial ring $Z_n[x]$ in the same way as we construct $\mathbb{F}_p[x]$. The next step is then to prove that n is a prime number if and only if $(x+1)^n = x^n + 1$ in $Z_n[x]$. Checking this condition for large n is too costly, and so a shortcut has to be found. The crucial idea is that if $(x+1)^n = x^n + 1$ holds in $Z_n[x]$, then also $(x+1)^n \equiv x^n + 1 \pmod{f(x)}$ for every polynomial $f(x) \in Z_n[x]$ of positive degree, where congruences in $Z_n[x]$ have the obvious meaning. For $f(x)$ we take $f(x) = (x+a)^h - 1$ with $a \in Z_n$ and $h \in \mathbb{N}$ suitably restricted so that we still get a correct primality test, but on the other hand a polynomial-time algorithm. In particular, the number of choices for a and h has to be at most of the order of magnitude $(\log n)^c$ for some constant $c > 0$. Mastering this balancing act is the beauty of the paper [1]. A detailed presentation of the AKS test starting from first principles is given in the recent book of Rempe-Gillen and Waldecker [167].

6.6.3 Primality tests for special numbers

The primality tests we have discussed so far can take any (odd) integer $n \geq 2$ as the input. It should not come as a surprise that if n has a very special form, then effective primality tests geared to the special nature of n can be designed. Specifically, we study the case where $n \pm 1$ is a power of 2. We start with $n = 2^s - 1$ for some integer $s \geq 2$.

Definition 6.6.13 A number of the form $2^s - 1$ with an integer $s \geq 2$ is called a *Mersenne number*. If $2^s - 1$ is a prime number, then it is called a *Mersenne prime*.

Example 6.6.14 The Mersenne numbers $2^2 - 1 = 3$, $2^3 - 1 = 7$, $2^5 - 1 = 31$, and $2^7 - 1 = 127$ are Mersenne primes. On the other hand, the Mersenne numbers $2^4 - 1 = 15$, $2^6 - 1 = 63$, and $2^{11} - 1 = 2047 = 23 \cdot 89$ are not Mersenne primes. Mersenne primes are of great practical value in pseudorandom number generation: the Mersenne prime $2^{31} - 1$ is a popular modulus in the linear congruential method (see Subsection 5.2.1) and the Mersenne prime $2^{19937} - 1$ plays an important role in the Mersenne twister (see Section 5.5).

You notice in Example 6.6.14 that in the cases where $2^s - 1$ is a Mersenne prime, the exponent s is always a prime number. Actually, it is a trivial general fact that $2^s - 1$ is a Mersenne prime only if s is a prime number (the converse does not hold: consider $2^{11} - 1$ in Example 6.6.14). Just note that if s has the nontrivial divisor $d \geq 2$, then $2^s - 1$ has the nontrivial divisor $2^d - 1$. Therefore we consider now only Mersenne numbers of the form $2^s - 1$ with a prime number s. Then the following astonishing criterion for primality can be established, where we omit the trivial case $s = 2$.

Theorem 6.6.15 *Let $n = 2^s - 1$ with a prime number $s \geq 3$. Then n is a prime number if and only if the sequence $u_0, u_1, \ldots$ of elements of Z_n defined recursively by*

$$u_0 = 4, \quad u_{k+1} \equiv u_k^2 - 2 \pmod{n} \quad \text{for } k = 0, 1, \ldots,$$

satisfies $u_{s-2} = 0$.

Proof. Let q be a prime factor of n and consider the polynomial

$$f(x) = x^2 - 2^{(s+1)/2}x - 1 \in \mathbb{F}_q[x]$$

with roots $\alpha, \beta \in \mathbb{F}_{q^2}$. Then

$$\alpha + \beta = 2^{(s+1)/2} \qquad \text{and} \qquad \alpha\beta = -1.$$

We view the u_k as elements of $\mathbb{F}_q$ and we show by induction that

$$u_k = \alpha^{2^{k+1}} + \beta^{2^{k+1}} \qquad \text{for all } k \geq 0. \tag{6.18}$$

This holds for $k = 0$ since $2^s \equiv 1 \pmod{q}$ and

$$\alpha^2 + \beta^2 = (\alpha + \beta)^2 - 2\alpha\beta = 2^{s+1} + 2 = 4 = u_0.$$

If (6.18) holds for some $k \geq 0$, then

$$\begin{aligned} u_{k+1} &= u_k^2 - 2 = (\alpha^{2^{k+1}} + \beta^{2^{k+1}})^2 - 2 = \alpha^{2^{k+2}} + \beta^{2^{k+2}} + 2(\alpha\beta)^{2^{k+1}} - 2 \\ &= \alpha^{2^{k+2}} + \beta^{2^{k+2}} + 2(-1)^{2^{k+1}} - 2 = \alpha^{2^{k+2}} + \beta^{2^{k+2}}, \end{aligned}$$

and the induction is complete.

If n is a prime number, then $q = n$. We have $\left(\frac{2}{n}\right) = 1$ since $\left(2^{(s+1)/2}\right)^2 \equiv 2^{s+1} \equiv 2 \pmod{n}$. From $n \equiv (-1)^s - 1 \equiv 1 \pmod{3}$ and $n \equiv 3 \pmod{4}$ we get $n \equiv 7 \pmod{12}$. Now 3 is a quadratic residue modulo a prime number $p \geq 5$ if and only if $p \equiv \pm 1 \pmod{12}$, as shown in [152, p. 139]. Hence $\left(\frac{3}{n}\right) = -1$, and then Proposition 1.2.24 yields $\left(\frac{6}{n}\right) = \left(\frac{2}{n}\right)\left(\frac{3}{n}\right) = -1$. The discriminant $\Delta(f)$ of the quadratic polynomial $f \in \mathbb{F}_n[x]$ is given by $\Delta(f) = 2^{s+1} + 4 = 6 \in \mathbb{F}_n$, which is a quadratic nonresidue modulo n. Therefore the usual formula for the roots of a quadratic polynomial shows that $\alpha, \beta \notin \mathbb{F}_n$, and so f is irreducible over $\mathbb{F}_n$. Hence $\alpha = \beta^n$ and $\beta = \alpha^n$ by Proposition 1.4.47, and thus

$$\alpha^{n+1} = \beta^{n+1} = \alpha\beta = -1.$$

Now (6.18) yields

$$-2 = \alpha^{n+1} + \beta^{n+1} = \alpha^{2^s} + \beta^{2^s} = u_{s-1} = u_{s-2}^2 - 2$$

in $\mathbb{F}_n$, and so $u_{s-2} = 0$.

Conversely, assume that $u_{s-2} = 0$ with a composite n and let q be any prime factor of n with $q^2 \leq n$. Then $\alpha^{2^{s-1}} + \beta^{2^{s-1}} = 0 \in \mathbb{F}_q$ by (6.18), and so $\alpha^{2^s} + (\alpha\beta)^{2^{s-1}} = 0 \in \mathbb{F}_q$. Thus $\alpha^{2^s} = -1 \in \mathbb{F}_q$ and $\alpha^{2^{s+1}} = 1 \in \mathbb{F}_q$. It follows that $\operatorname{ord}(\alpha) = 2^{s+1}$ in the multiplicative group $\mathbb{F}_{q^2}^*$, and so 2^{s+1} divides $q^2 - 1$ by Proposition 1.3.11. But this is impossible since $q^2 - 1 < n < 2^{s+1}$. □

The primality test for Mersenne numbers based on Theorem 6.6.15 is called the *Lucas-Lehmer test*. Since $s = \log_2(n+1)$, it is a deterministic polynomial-time algorithm.

Example 6.6.16 Just for illustration, consider the toy example $n = 2^s - 1$ with $s = 7$. The numbers u_k, $k = 0, 1, \ldots, 5$, from Theorem 6.6.15 are computed in the following table.

k	0	1	2	3	4	5
u_k	4	14	67	42	111	0

Since $u_5 = 0$, we infer from Theorem 6.6.15 that n is a prime number. This can also be verified directly since $n = 2^7 - 1 = 127$.

For several centuries there is a competition about explicitly finding larger and larger prime numbers. For a long time now, the new world records are Mersenne primes since for them we have very efficient deterministic primality tests such as the Lucas-Lehmer test. In the media a new world record is sometimes reported as "such and such is the largest prime number", which is of course nonsense since there are infinitely many prime numbers. The gripping story of the quest for large prime numbers is told at length in [6, Section 1.2].

Now we consider numbers of the form 2^s+1 with an integer $s \geq 1$. If d is a divisor of s and $s/d \geq 3$ is odd, then 2^d+1 is a nontrivial divisor of 2^s+1. Therefore 2^s+1 is a prime number only if s is a power of 2.

Definition 6.6.17 A number N_k of the form $N_k = 2^{2^k}+1$ with an integer $k \geq 0$ is called a *Fermat number*, and it is called a *Fermat prime* if it is a prime number.

Example 6.6.18 The first five Fermat numbers $N_0 = 2^1+1 = 3$, $N_1 = 2^2+1 = 5$, $N_2 = 2^4+1 = 17$, $N_3 = 2^8+1 = 257$, and $N_4 = 2^{16}+1 = 65537$ are Fermat primes, which led Fermat to conjecture that all Fermat numbers are prime numbers. It caused quite a stir in the 18th century when Euler discovered the nontrivial prime factor 641 of the Fermat number $N_5 = 2^{32}+1$, thus demolishing Fermat's conjecture.

Some people believe that there are no Fermat primes beyond $2^{16}+1$, and indeed none have been found so far. The following is an easy criterion for Fermat primes.

Theorem 6.6.19 *The Fermat number $N_k = 2^{2^k}+1$ with $k \geq 1$ is a prime number if and only if*

$$3^{(N_k-1)/2} \equiv -1 \pmod{N_k}. \tag{6.19}$$

Proof. We note that $N_k \equiv (-1)^{2^k}+1 \equiv 2 \pmod 3$ and $N_k \equiv 1 \pmod 4$ for $k \geq 1$, and so $N_k \equiv 5 \pmod{12}$. If N_k is a prime number, then the criterion for the quadratic-residue behavior of 3 mentioned in the proof of Theorem 6.6.15 shows that 3 is a quadratic nonresidue modulo N_k. Hence (6.19) follows from Proposition 1.2.23.

Conversely, if (6.19) holds for some $k \geq 1$, then $3^{N_k-1} \equiv 1 \pmod{N_k}$. Since $N_k - 1$ is a power of 2, we infer that 3 has order $N_k - 1$ in the multiplicative group R_{N_k} of order $|R_{N_k}| = \phi(N_k) \leq N_k - 1$. Hence $\phi(N_k) = N_k - 1$, which implies that N_k is a prime number. □

Remark 6.6.20 It is a curious fact that Euclid's theorem on the infinitude of prime numbers (see Theorem 1.1.12) can be proved by means of Fermat numbers. From $N_{k+1} = (N_k-1)^2+1 = N_k(N_k-2)+2$ for all $k \geq 0$, we derive by induction that $N_{k+1} = N_0 \cdots N_k + 2$ for all $k \geq 0$. We claim that $\gcd(N_k, N_m) = 1$ whenever $0 \leq k < m$. For if d is a positive common divisor of N_k and N_m, then d divides both $N_0 \cdots N_{m-1}$ and N_m, and so d divides $N_m - N_0 \cdots N_{m-1} = 2$. But all Fermat numbers are odd, and so $d = 1$. Thus, if for each $k = 0, 1, \ldots$ we choose a prime factor p_k of N_k, then we get an infinite sequence $p_0, p_1, \ldots$ of distinct prime numbers.

6.7 Two more applications

6.7.1 Benford's law

We could go on and on with applications of number theory, such is the richness of the subject, but who would read a textbook with over 1000 pages? So it is time to reach an end, but nevertheless we cannot refrain from picking two more raisins (we hope tasty ones) from the cake. We start with a discussion of digit distributions, and in the next subsection we present an application of number theory to raster graphics.

Is it one of your favorite recreational activities to read long lists of bookkeeping data? If so, then you may have noticed the rather skew distribution of the leading digits in these data. Some attentive fellows observed this as an empirical fact, and the physicist Frank Benford is credited for this discovery since he carried out a wide-ranging study of data in the 1930s. This phenomenon, which is known as Benford's law or the first-digit law, occurs not only in accounting data, but also in large collections of physical and mathematical constants, of stock prices, of geographic data like lengths of rivers, and so on.

After rounding off and suitably scaling the data, we can assume that we are talking about the leading digits in a sequence of positive integers. In the binary case, we have the extreme situation where the leading digit in the binary representation of a positive integer is always 1. In the decimal case considered by Benford, he noticed that the asymptotic proportion of 1 as a leading digit is equal to $\log_{10} 2$ (around 30.1%), the asymptotic proportion of 2 as a leading digit is equal to $\log_{10} \frac{3}{2}$ (around 17.6%), and so on until the asymptotic proportion of 9 as a leading digit which is $\log_{10} \frac{10}{9}$ (around 4.6%). For every integer $b \geq 2$ and every positive integer k, we denote by $\ell_b(k)$ the leading digit in the digit expansion of k in base b. For instance, $\ell_{10}(37) = 3$ and $\ell_{10}(143) = 1$.

Definition 6.7.1 For an integer $b \geq 2$, a sequence $(k_n)_{n=1}^{\infty}$ of positive integers satisfies *Benford's law* (or the *first-digit law*) in the base b if

$$\lim_{N\to\infty} \frac{\#\{1 \leq n \leq N : \ell_b(k_n) = d\}}{N} = \log_b \left(1 + \frac{1}{d}\right) \qquad \text{for } d = 1, \ldots, b-1,$$

where $\log_b$ denotes the logarithm to the base b.

There is an important sufficient condition for Benford's law which connects this law with the theory of uniformly distributed sequences in Subsection 4.1.1.

Theorem 6.7.2 *Let $b \geq 2$ be an integer and let $(k_n)_{n=1}^{\infty}$ be a sequence of positive integers. If the sequence $(\log_b k_n)_{n=1}^{\infty}$ is uniformly distributed modulo 1, then the sequence $(k_n)_{n=1}^{\infty}$ satisfies Benford's law in the base b.*

Proof. We fix the base $b \geq 2$. Then for $k \in \mathbb{N}$ and $d \in \{1, \ldots, b-1\}$, we have $\ell_b(k) = d$ if and only if $db^m \leq k < (d+1)b^m$ for some integer $m \geq 0$. By taking logarithms, we obtain the equivalent condition $m + \log_b d \leq \log_b k < m + \log_b(d+1)$. This is the same as saying that the fractional part $\{\log_b k\}$ satisfies

$$\{\log_b k\} \in [\log_b d, \log_b(d+1)).$$

Therefore

$$\#\{1 \leq n \leq N : \ell_b(k_n) = d\} = \#\{1 \leq n \leq N : \{\log_b k_n\} \in [\log_b d, \log_b(d+1))\}$$

for all integers $N \geq 1$ and all $d \in \{1, \ldots, b-1\}$. The desired result follows now from Definition 4.1.8 and Theorem 4.1.6. □

Example 6.7.3 For every integer $a \geq 2$ which is not a power of 10, the sequence $(a^n)_{n=1}^{\infty}$ of powers of a satisfies Benford's law in the base 10. This is a simple consequence of Theorem 6.7.2. Note that $\log_{10} a^n = n \log_{10} a$ for all $n \geq 1$. The number $\log_{10} a$ is irrational, for if we had $\log_{10} a = r/s$ with $r, s \in \mathbb{N}$, then $a^s = 10^r$, which is impossible under the given condition on a. Therefore the sequence $(\log_{10} a^n)_{n=1}^{\infty}$ is uniformly distributed modulo 1 by Theorem 4.1.10. Analogous examples can be constructed with 10 replaced by an arbitrary base $b \geq 2$.

Example 6.7.4 It takes a bit more work to prove that the sequence $(F_n)_{n=1}^{\infty}$ of Fibonacci numbers, which is defined recursively by $F_1 = F_2 = 1$ and $F_{n+2} = F_{n+1} + F_n$ for $n = 1, 2, \ldots$, satisfies Benford's law in every base $b \geq 2$. First we show by straightforward induction that

$$F_n = \frac{1}{\sqrt{5}}(\alpha^n - \beta^n) \qquad \text{for } n = 1, 2, \ldots, \tag{6.20}$$

where $\alpha = (1+\sqrt{5})/2$ and $\beta = (1-\sqrt{5})/2$. Now we put $x_n = n \log_b \alpha$ for all $n \geq 1$. If we assume for the moment that $\log_b \alpha$ is irrational, then the sequence $(x_n)_{n=1}^{\infty}$ is uniformly distributed modulo 1 by Theorem 4.1.10. Furthermore, (6.20) implies that

$$\lim_{n\to\infty} (\log_b F_n - x_n) = \lim_{n\to\infty} \log_b \frac{1 - (\beta/\alpha)^n}{\sqrt{5}} = -\log_b \sqrt{5}.$$

Then an easy application of Theorem 4.1.9 shows that the sequence $(\log_b F_n)_{n=1}^{\infty}$ is uniformly distributed modulo 1, and so $(F_n)_{n=1}^{\infty}$ satisfies Benford's law by Theorem 6.7.2. It remains to prove that $\log_b \alpha$ is irrational. If we had $\log_b \alpha = r/s$ with $r, s \in \mathbb{N}$, then $(1+\sqrt{5})^s = 2^s b^r$. Now by induction $(1+\sqrt{5})^s = a_s + c_s\sqrt{5}$ with $a_s, c_s \in \mathbb{N}$, hence $c_s\sqrt{5} = 2^s b^r - a_s$, an obvious contradiction to the fact that $\sqrt{5}$ is irrational. More general linear recurring sequences satisfying Benford's law are constructed in the paper [120]. It is shown in [92] that the sequence $(n!)_{n=1}^{\infty}$ of factorials satisfies Benford's law in the base 10, and the proof is again founded on Theorem 6.7.2.

Example 6.7.5 Here is an interesting negative example. Let $(n)_{n=1}^{\infty}$ be the sequence of positive integers in their natural order and consider the standard decimal case $b = 10$. Among the first 20 terms of this sequence, 11 have leading digit 1, among the first 200 terms of this sequence, 111 have leading digit 1, and in general among the first $2 \cdot 10^r$ terms of this sequence with $r \in \mathbb{N}$, there are $\sum_{i=0}^{r} 10^i = (10^{r+1} - 1)/9$ numbers with leading digit 1. Therefore

$$\lim_{r \to \infty} \frac{\#\{1 \leq n \leq 2 \cdot 10^r : \ell_{10}(n) = 1\}}{2 \cdot 10^r} = \lim_{r \to \infty} \frac{10^{r+1} - 1}{18 \cdot 10^r} = \frac{5}{9}.$$

But $\frac{5}{9} = 0.555\ldots > \log_{10} 2 = 0.301\ldots$, and so the sequence $(n)_{n=1}^{\infty}$ does not satisfy Benford's law in the base 10.

The recent book of Kossovsky [88] is a treasure trove for the history and the applications of Benford's law. As an example of an application, we mention that Benford's law can be utilized in fraud detection since deceitfully concocted data may deviate from Benford's law. Cheaters tend to use the uniform distribution of leading digits rather than the distribution in Definition 6.7.1.

6.7.2 An application to raster graphics

Let us come back to the Fibonacci numbers in Example 6.7.4. An intriguing application of Fibonacci numbers was discovered by Chor, Leiserson, and Rivest [23] (see also [24]). By the way, Rivest is also the R in the RSA public-key cryptosystem (see Subsection 2.3.2). The approach and the proof of the main result in [23] can be considerably simplified, as we show in the following.

The problem addressed in [23] is highly relevant for the efficient operation of computer screens, TV screens, and videos. Let us talk simply about screens to avoid cumbersome language. The pixels of a screen are controlled by random-access memory cells, but in a high-resolution screen there are of course many more pixels than memory cells. This raises the question of how to assign pixels to memory cells so that large areas of the screen can be updated simultaneously. It is convenient to label the rows and columns of pixels on the screen by nonnegative integers. If there are M memory cells, then we label the memory cells by the elements of $Z_M = \{0, 1, \ldots, M-1\}$, the least residue system modulo M.

Figure 6.1 shows the standard assignment of pixels to memory cells. In row 0 the first M pixels are assigned to the memory cells $0, 1, \ldots, M-1$ in that order, and this assignment is repeated periodically with period length M. The same pattern is used in all the other rows. The standard assignment is very efficient for rowwise updating since any M consecutive pixels in any row are assigned to different memory cells and can therefore be updated simultaneously. On the other hand, columnwise updating is a stumbling block since all pixels in a given column are assigned to the

$$\begin{array}{ccccccccccccc}
0 & 1 & 2 & 3 & \ldots & M-1 & 0 & 1 & 2 & 3 & \ldots & M-1 & \ldots \\
0 & 1 & 2 & 3 & \ldots & M-1 & 0 & 1 & 2 & 3 & \ldots & M-1 & \ldots \\
0 & 1 & 2 & 3 & \ldots & M-1 & 0 & 1 & 2 & 3 & \ldots & M-1 & \ldots \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \\
0 & 1 & 2 & 3 & \ldots & M-1 & 0 & 1 & 2 & 3 & \ldots & M-1 & \ldots
\end{array}$$

Figure 6.1: The standard assignment of pixels to memory cells

same memory cell, and so these pixels can be updated only one after the other and parallelization is not possible.

There should be a better organization of raster graphics than the standard assignment, and this is what the work of Chor, Leiserson, and Rivest [23] is all about. The aim is to find an assignment of pixels to memory cells such that the pixels in all rectangles of limited area can be updated simultaneously, that is, the labels of the assigned memory cells in any such rectangle are different. When we speak of a rectangle, we mean a rectangle with horizontal and vertical sides (that is, no tilted rectangles are considered) and with positive integers as side lengths (on the scale of the pixels), and the area of such a rectangle is defined to be the number of pixels in the rectangle.

We describe the construction in [23] in an explicit and simplified form. For a fixed integer $n \geq 2$, let $M = F_{2n+1}$ be the Fibonacci number with index $2n+1$. For all $r, s = 0, 1, \ldots,$ the pixel in row r and column s is assigned to the memory cell with the label $a(r,s) \in Z_M$ that is uniquely determined by the congruence

$$a(r,s) \equiv F_{2n}r + s \pmod{M}, \tag{6.21}$$

where F_{2n} is the Fibonacci number with index $2n$. We may call this the *Fibonacci assignment*. Row 0 of the Fibonacci assignment is identical with row 0 of the standard assignment for the same number M of memory cells. The other rows of the Fibonacci assignment are cyclic shifts of row 0, with the shift parameter in row r given by the least residue of $F_{2n}r$ modulo M. Figure 6.2 shows the Fibonacci assignment for $n = 3$, that is, with $M = F_7 = 13$. It follows from Theorem 6.7.6 below that in this case, the memory-cell labels in every rectangle of area at most 11 are different. A few rectangles of this type are highlighted in the figure.

Theorem 6.7.6 *For an integer $n \geq 2$, put $M = F_{2n+1}$ and $N = F_nF_{n+1} + F_{n+2}$. Then the Fibonacci assignment with M memory cells has the property that in every rectangle of area at most N the memory-cell labels are different.*

Proof. The proof uses the theory of continued fractions (see Subsection 4.2.1) and

0	1	2	3	4	5	6	7	8	9	10	11	12	0	1	2	...
8	9	10	**11**	**12**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	9	10	...
3	4	5	6	7	8	9	10	11	12	0	1	2	3	4	5	...
11	12	0	1	2	3	4	5	6	7	8	9	10	11	12	0	...
6	7	8	**9**	**10**	**11**	**12**	**0**	1	2	3	4	5	6	7	8	...
1	2	3	**4**	**5**	**6**	**7**	**8**	9	10	**11**	**12**	0	1	2	3	...
9	10	11	12	0	1	2	3	4	5	**6**	**7**	8	9	10	11	...
4	5	6	7	8	9	10	11	12	0	**1**	**2**	3	4	5	6	...
12	0	1	2	3	4	5	6	7	8	**9**	**10**	11	12	0	1	...
7	8	9	10	11	12	0	1	2	3	**4**	**5**	6	7	8	9	...
2	3	4	5	6	7	8	9	10	11	12	0	1	2	3	4	...
10	11	12	0	**1**	**2**	**3**	4	5	6	7	8	9	10	11	12	...
5	6	7	8	**9**	**10**	**11**	12	0	1	2	3	4	5	6	7	...
0	1	2	3	**4**	**5**	**6**	7	8	9	10	11	12	0	1	2	...
8	9	10	11	12	0	1	2	3	4	5	6	7	8	9	10	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

Figure 6.2: The Fibonacci assignment for $M = F_7 = 13$

identities for Fibonacci numbers. First we show that $N \le M$. Indeed,

$$\begin{aligned} N &= F_n F_{n+1} + F_{n+1} + F_n \le F_n F_{n+1} + F_{n+1} F_{n-1} + F_n^2 \\ &= F_{n+1}(F_n + F_{n-1}) + F_n^2 = F_{n+1}^2 + F_n^2 = M, \end{aligned}$$

where the last identity is obtained from (6.20). Now we take an arbitrary $R \times S$ rectangle of area $RS \le N$ and we assume that two memory-cell labels in this rectangle are equal. This means that $a(r_1, s_1) = a(r_2, s_2)$ for some $l \le r_1, r_2 \le l + R - 1$ and $m \le s_1, s_2 \le m + S - 1$ and with some integers $l, m \ge 0$. We have to show that $r_1 = r_2$ and $s_1 = s_2$. Note that (6.21) yields $F_{2n}(r_1 - r_2) \equiv s_2 - s_1 \pmod{M}$. If $r_1 = r_2$, then $s_1 \equiv s_2 \pmod{M}$ and $|s_1 - s_2| \le S - 1 < N \le M$. Therefore $s_1 = s_2$ and we are done.

Thus, we can assume by way of contradiction that $r_1 > r_2$. Putting $h = r_1 - r_2$ and $j = s_2 - s_1$, we obtain

$$F_{2n} h \equiv j \pmod{M} \tag{6.22}$$

with $1 \le h \le R - 1$ and $|j| \le S - 1$. In particular $1 \le h < N \le M = F_{2n+1}$, and so there exists an integer i with $2 \le i \le 2n$ such that $F_i \le h < F_{i+1}$. Now we use the fact noted in Example 4.3.15 that the rational number F_{2n}/F_{2n+1} has the finite continued fraction expansion

$$\frac{F_{2n}}{F_{2n+1}} = [0; \underbrace{1, 1, \ldots, 1}_{2n}],$$

and if we terminate this expansion after the kth entry 1 with $k \le 2n$, then we get the continued fraction expansion of F_k/F_{k+1}. Therefore [152, Theorem 7.13] shows that for every $t \in \mathbb{Z}$ the inequality

$$\left|\frac{F_{2n}}{F_{2n+1}}h - t\right| \geq \left|\frac{F_{2n}}{F_{2n+1}}F_i - F_{i-1}\right|$$

is valid. According to (6.22), we can choose $t \in \mathbb{Z}$ such that $j = F_{2n}h - F_{2n+1}t$, and so

$$|j| \geq |F_{2n}F_i - F_{2n+1}F_{i-1}| = F_{2n+1-i},$$

where the last identity is again obtained from (6.20); see also Exercise 6.40. It follows that

$$N \geq RS \geq (h+1)(|j|+1) \geq (F_i+1)(F_{2n+1-i}+1) := G(i,n).$$

Note that $G(i,n)$ makes sense also for $i = 1$. If we can show that

$$\min_{1 \leq i \leq 2n} G(i,n) = G(n,n), \tag{6.23}$$

then we arrive at the desired contradiction since $G(n,n) = (F_n+1)(F_{n+1}+1) = N+1$.

In view of $G(i,n) = G(2n+1-i,n)$ for $1 \leq i \leq 2n$, we can guarantee (6.23) by proving that $G(i,n) \geq G(i+1,n)$ for $1 \leq i \leq n-1$. Using another Fibonacci identity obtained from (6.20) (see again Exercise 6.40), we get

$$G(i,n) - G(i+1,n) = (-1)^{i+1}F_{2n-2i} + F_{2n-i-1} - F_{i-1},$$

with the understanding that $F_0 = 0$. If i is odd, then $G(i,n) \geq G(i+1,n)$ follows from $F_{2n-i-1} \geq F_{i-1}$. If i is even, then

$$G(i,n) - G(i+1,n) = F_{2n-i-2} + F_{2n-i-3} - F_{2n-2i} - F_{i-1},$$

and also $F_{2n-i-2} \geq F_{2n-2i}$ and $F_{2n-i-3} \geq F_{i-1}$ since $2 \leq i \leq n-1$. Therefore (6.23) is shown. □

It is clear from the proof above that with minor modifications we can deal also with the case where $M = F_{2n}$ with $n \geq 2$. Here is a small table of the corresponding pairs of numbers M and N obtained from Theorem 6.7.6 with $M \leq 1000$.

M	5	13	34	89	233	610
N	5	11	23	53	125	307

As we said at the beginning of this section, there are many more applications of number theory, but most of them are easy and require only elementary number theory. Just to whet your appetite, we mention applications to visibility problems [3], [183], [184], to fast convolution algorithms [31], [153], to binary search trees [37], to cable splicing [155, Section 12.8], to music theory [166], and to card tricks [118, p. 632], [175]. The conference volumes [16] and [122] contain attractive selections of applications of number theory that are out of the common. You may want to explore some of these applications at your leisure.

Exercises

6.1 Consider the check-digit system over Z_9 defined by the control equation

$$\sum_{i=1}^{12} a_i \equiv 8 \pmod 9.$$

The serial numbers of Euro banknotes, with a proper interpretation of letters, are based on this check-digit system.

(a) Verify that 402387040034 satisfies the control equation.

(b) Show that this check-digit system detects neither neighbor transpositions nor jump transpositions.

(c) Show that this check-digit system detects both twin errors and jump twin errors.

6.2 (a) Show that the self-map of $\mathbb{F}_7$ defined by the polynomial $f(x) = x^4 + 3x \in \mathbb{F}_7[x]$ is a complete mapping of $\mathbb{F}_7$.

(b) Show that the self-map of $\mathbb{F}_{11}$ defined by the polynomial $f(x) = 2x^6 + 7x \in \mathbb{F}_{11}[x]$ is a complete mapping of $\mathbb{F}_{11}$.

6.3 Prove that if f is a complete mapping of the finite abelian group G, then the inverse map f^{-1} is also a complete mapping of G.

6.4 Let $q \geq 5$ be a power of an odd prime. Prove that if the self-map of $\mathbb{F}_q$ defined by the polynomial $f \in \mathbb{F}_q[x]$ with $\deg(f) < q$ is a complete mapping of $\mathbb{F}_q$, then necessarily $\deg(f) \leq q - 3$. (Hint: see [140].)

6.5 Show that for every odd prime number p, the set $\{0, 1, \ldots, (p-1)/2\}$ is a $(1, 1; p)$-covering set of minimal size.

6.6 Show that for every odd prime number p, the set $\{1, 2, \ldots, (p-1)/2\}$ is a $(1, 1; p)$-packing set of maximal size.

6.7 Let p be an odd prime number. For a positive divisor k of $\frac{p-1}{2}$ and $s \in \mathbb{N}$, consider the set

$$A_{k,s} = \{a_1^k + a_1^{-k} + \cdots + a_s^k + a_s^{-k} \in \mathbb{F}_p : a_1, \ldots, a_s \in \mathbb{F}_p^*\}.$$

(a) Show that $|A_{k,1}| = \frac{p-1}{2k} + 1$.

(b) Show that $A_{k,2k} = \mathbb{F}_p$.

(c) Verify that $(a^k + a^{-k})(b^k + b^{-k}) = (ab)^k + (ab)^{-k} + (ab^{-1})^k + (ab^{-1})^{-k}$ for all $a, b \in \mathbb{F}_p^*$.

(d) Show that $A_{k,16} = \mathbb{F}_p$ if $p \geq 8k^2$.

6.8 Let C be a nontrivial linear code. Prove that the covering radius $\rho(C)$ is equal to the maximum Hamming weight of all coset leaders.

6.9 Suppose that the code C_1 is a proper subset of the code C_2. Prove that $\rho(C_1) \geq d(C_2)$.

6.10 (a) Construct a Hadamard matrix of order 8 in two different ways.

(b) Construct a Hadamard matrix of order 12 by a method that is different from the one in Example 6.4.10.

6.11 If $H = (h_{ij})_{1 \leq i,j \leq m}$ is an $m \times m$ matrix and K an $n \times n$ matrix over $\mathbb{R}$, then the Kronecker product $H \otimes K$ is the $(mn) \times (mn)$ matrix given by

$$H \otimes K = \begin{pmatrix} h_{11}K & h_{12}K & \dots & h_{1m}K \\ h_{21}K & h_{22}K & \dots & h_{2m}K \\ \vdots & \vdots & & \vdots \\ h_{m1}K & h_{m2}K & \dots & h_{mm}K \end{pmatrix}.$$

Prove that if H and K are Hadamard matrices, then $H \otimes K$ is a Hadamard matrix. Thus, whenever Hadamard matrices of orders m and n exist, then there exists a Hadamard matrix of order mn.

6.12 An $n \times n$ matrix M over $\mathbb{R}$ with $n \geq 2$ for which the entries on the main diagonal are 0, all other entries are 1 or -1, and which satisfies $MM^\top = (n-1)E_n$ is called a *conference matrix* of order n. Prove that the matrix J in the proof of Theorem 6.4.8 is a conference matrix of order $q+1$.

6.13 Prove that if a conference matrix of order n exists, then n must be even.

6.14 Let $A(n,d)$ be as in Definition 6.4.11. Prove that $A(n,1) = 2^n$ and $A(n,n) = 2$ for all integers $n \geq 1$.

6.15 Prove that $A(n,d) \geq 2^n / \big(\sum_{i=0}^{d-1} \binom{n}{i} \big)$ for $1 \leq d \leq n$.

6.16 Prove that $A(n,d) \leq 2^n / \big(\sum_{i=0}^{\lfloor (d-1)/2 \rfloor} \binom{n}{i} \big)$ for $1 \leq d \leq n$.

6.17 Prove that $A(n,d) \leq 2^{n-d+1}$ for $1 \leq d \leq n$.

6.18 Show in detail that the Boolean function g in the proof of Theorem 6.4.27 is bent.

6.19 Show in detail that in the proof of Theorem 6.4.27 we have indeed $F(\mathbf{u}) = 2^{-n/2}\hat{G}(\mathbf{u})$ for all $\mathbf{u} \in \mathbb{F}_2^n$.

6.20 Show that a Boolean function f of n variables is bent if and only if, for all nonzero vectors $\mathbf{v} \in \mathbb{F}_2^n$, the Boolean function $f_{\mathbf{v}}$ defined by $f_{\mathbf{v}}(\mathbf{u}) = f(\mathbf{u} + \mathbf{v}) + f(\mathbf{u})$ for all $\mathbf{u} \in \mathbb{F}_2^n$ attains the values 0 and 1 equally often.

6.21 Let $f = f(\mathbf{u})$ be a bent function of m variables and let $g = g(\mathbf{v})$ be a bent function of n variables. Prove that $h(\mathbf{u}, \mathbf{v}) = f(\mathbf{u}) + g(\mathbf{v})$ is a bent function of $m + n$ variables.

6.22 Show that the ANF exists for every Boolean function and is unique.

6.23 Determine the ANF and the algebraic degree of the Boolean function $f(x_1, x_2, x_3)$ given by the following table.

x_1	x_2	x_3	$f(x_1, x_2, x_3)$
0	0	0	0
0	0	1	1
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	1

6.24 Determine the nonlinearity of the Boolean functions $f_1(x_1, x_2) = x_1 + x_2 + 1$, $f_2(x_1, x_2) = x_1x_2$, and $f_3(x_1, x_2, x_3, x_4) = x_1x_2 + x_3x_4$.

6.25 Let $n \in \mathbb{N}$ be even and assume that the Boolean function f of n variables satisfies $\mathcal{NL}(f) = 2^{n-1} - 2^{n/2-1}$. Show that $\deg(f) = 2$ if $n = 2$ and $\deg(f) \leq n/2$ if $n \geq 4$.

6.26 Show that 2 is a quadratic residue modulo the odd prime number p if and only if $p \equiv \pm 1 \pmod 8$.

6.27 Determine the exact value of the algebraic degree and of the nonlinearity of the Boolean function g defined by (6.17) for $p \in \{5, 7, 11\}$.

6.28 Factor the number 91 using the Shor algorithm. (Show first that the multiplicative order of 4 modulo 91 is 6 and that $4^3 \not\equiv -1 \pmod{91}$.)

6.29 Prove that every Carmichael number is odd.

6.30 Prove Wilson's theorem that an integer $n \geq 2$ is a prime number if and only if $(n-1)! \equiv -1 \pmod n$.

6.31 For all odd integers $n \geq 3$ and all $a, b \in \mathbb{Z}$, prove that

$$\left(\frac{ab}{n}\right) = \left(\frac{a}{n}\right)\left(\frac{b}{n}\right).$$

6.32 For all odd integers $m, n \geq 3$ and all $a \in \mathbb{Z}$, prove that

$$\left(\frac{a}{mn}\right) = \left(\frac{a}{m}\right)\left(\frac{a}{n}\right).$$

6.33 For all odd integers $n \geq 3$, prove that

$$\left(\frac{-1}{n}\right) = (-1)^{(n-1)/2}.$$

6.34 Show that $2^{11} - 1$ is not a Mersenne prime by verifying that $2^{11} \equiv 1 \pmod{23}$.

6.35 Prove that the Fermat number $N_k = 2^{2^k} + 1$ with $k \geq 2$ is a prime number if and only if $5^{(N_k-1)/2} \equiv -1 \pmod{N_k}$. (Hint: use that 5 is a quadratic residue modulo the odd prime number p if and only if $p \equiv \pm 1 \pmod 5$, according to [152, Theorem 3.5].)

6.36 Let p be an odd prime number. Prove that p is a Fermat prime if and only if every quadratic nonresidue modulo p is a primitive root modulo p.

6.37 Let $a \geq 2$ and $b \geq 2$ be integers. Prove that $\log_b a$ is a rational number, say $\log_b a = r/s$ with $r, s \in \mathbb{N}$ and $\gcd(r, s) = 1$, if and only if there exists an integer $c \geq 2$ such that $a = c^r$ and $b = c^s$. This criterion yields a generalization of Example 6.7.3.

6.38 Prove that the sequence $(L_n)_{n=1}^{\infty}$ of integers defined by $L_1 = L_2 = 1$ and $L_{n+2} = 2L_{n+1} + L_n$ for $n = 1, 2, \ldots$ satisfies Benford's law in every base $b \geq 2$.

6.39 Prove the Fibonacci identity $F_{n+1}^2 + F_n^2 = F_{2n+1}$ for every integer $n \geq 1$. This identity was used in the proof of Theorem 6.7.6.

6.40 Prove the Fibonacci identity $F_i F_j - F_{i+1} F_{j-1} = (-1)^{j-1} F_{i-j+1}$ for all integers $i \geq j \geq 1$, where we put $F_0 = 0$. Special cases of this identity were used in the proof of Theorem 6.7.6.

Bibliography

[1] M. Agrawal, N. Kayal, and N. Saxena, PRIMES is in P, *Annals of Math.* (2) **160**, 781–793 (2004).

[2] W.R. Alford, A. Granville, and C. Pomerance, There are infinitely many Carmichael numbers, *Annals of Math.* (2) **139**, 703–722 (1994).

[3] T.T. Allen, Polya's orchard problem, *Amer. Math. Monthly* **93**, 98–104 (1986).

[4] N. Alon, M.B. Nathanson, and I. Ruzsa, The polynomial method and restricted sums of congruence classes, *J. Number Theory* **56**, 404–417 (1996).

[5] N. Aydin and T. Asamov, Search for good linear codes in the class of quasi-cyclic and related codes, *Selected Topics in Information and Coding Theory* (I. Woungang, S. Misra, and S.C. Misra, eds.), pp. 239–285, World Scientific, Singapore, 2010.

[6] E. Bach and J. Shallit, *Algorithmic Number Theory*, Vol. 1: *Efficient Algorithms*, MIT Press, Cambridge, MA, 1996.

[7] N.S. Bakhvalov, Approximate computation of multiple integrals, *Vestnik Moskov. Univ. Ser. Mat. Mekh. Astr. Fiz. Khim.* **1959**, no. 4, 3–18 (1959). (Russian)

[8] M. Baldi, *QC-LDPC Code-Based Cryptography*, Springer, Berlin, 2014.

[9] J. Beck, Probabilistic Diophantine approximation. I. Kronecker sequences, *Annals of Math.* (2) **140**, 449–502 (1994).

[10] E.R. Berlekamp, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.

[11] D.J. Bernstein, J. Buchmann, and E. Dahmen (eds.), *Post-Quantum Cryptography*, Springer, Berlin, 2009.

[12] I.F. Blake, G. Seroussi, and N.P. Smart, *Elliptic Curves in Cryptography*, Cambridge University Press, Cambridge, 1999.

[13] R.C. Bose and D.K. Ray-Chaudhuri, On a class of error correcting binary group codes, *Information and Control* **3**, 68–79 (1960).

[14] H. Brass and K. Petras, *Quadrature Theory: The Theory of Numerical Integration on a Compact Interval*, American Mathematical Society, Providence, RI, 2011.

[15] J.A. Buchmann, *Introduction to Cryptography*, Springer, New York, 2001.

[16] S.A. Burr (ed.), *The Unreasonable Effectiveness of Number Theory*, Proc. Symposia in Applied Mathematics, Vol. **46**, American Mathematical Society, Providence, RI, 1992.

[17] K.A. Bush, Orthogonal arrays of index unity, *Ann. Math. Statistics* **23**, 426–434 (1952).

[18] C. Carlet, Boolean functions for cryptography and error correcting codes, *Boolean Models and Methods in Mathematics, Computer Science, and Engineering* (Y. Crama and P.L. Hammer, eds.), pp. 257–397, Cambridge University Press, Cambridge, 2010.

[19] C. Carlet, Vectorial Boolean functions for cryptography, *Boolean Models and Methods in Mathematics, Computer Science, and Engineering* (Y. Crama and P.L. Hammer, eds.), pp. 398–469, Cambridge University Press, Cambridge, 2010.

[20] F. Castro and I. Rubio, Diagonal equations, *Handbook of Finite Fields* (G.L. Mullen and D. Panario, eds.), pp. 206–213, CRC Press, Boca Raton, FL, 2013.

[21] D.G. Champernowne, The construction of decimals normal in the scale of ten, *J. London Math. Soc.* **8**, 254–260 (1933).

[22] Z.X. Chen, I.E. Shparlinski, and A. Winterhof, Covering sets for limited-magnitude errors, *IEEE Trans. Inform. Theory* **60**, 5315–5321 (2014).

[23] B. Chor, C.E. Leiserson, and R.L. Rivest, An application of number theory to the organization of raster-graphics memory, *Proc. 23rd Symposium on Foundations of Computer Science* (Chicago, 1982), pp. 92–99, IEEE Computer Society, Los Angeles, 1982.

[24] B. Chor, C.E. Leiserson, R.L. Rivest, and J.B. Shearer, An application of number theory to the organization of raster-graphics memory, *J. Assoc. Computing Machinery* **33**, 86–104 (1986).

[25] K.L. Chung, An estimate concerning the Kolmogoroff limit distribution, *Trans. Amer. Math. Soc.* **67**, 36–50 (1949).

[26] G. Cohen, I. Honkala, S. Litsyn, and A. Lobstein, *Covering Codes*, North-Holland, Amsterdam, 1997.

[27] H. Cohen, G. Frey, R. Avanzi, C. Doche, T. Lange, K. Nguyen, and F. Vercauteren (eds.), *Handbook of Elliptic and Hyperelliptic Curve Cryptography*, CRC Press, Boca Raton, FL, 2006.

[28] J.G. van der Corput, Verteilungsfunktionen I, II, *Nederl. Akad. Wetensch. Proc. Ser. B* **38**, 813–821, 1058–1066 (1935). (German)

[29] R.R. Coveyou, Random number generation is too important to be left to chance, Studies in Applied Mathematics, Vol. **3**, pp. 70–111, SIAM, Philadelphia, 1969.

[30] R. Crandall and C. Pomerance, *Prime Numbers: A Computational Perspective*, Springer, New York, 2001.

[31] R. Creutzburg and M. Tasche, Number-theoretic transforms of prescribed length, *Math. Comp.* **47**, 693–701 (1986).

[32] T.W. Cusick, C. Ding, and A. Renvall, *Stream Ciphers and Number Theory*, Elsevier, Amsterdam, 1998.

[33] T.W. Cusick and P. Stănică, *Cryptographic Boolean Functions and Applications*, Elsevier and Academic Press, Amsterdam, 2009.

[34] J. Daemen and V. Rijmen, *The Design of Rijndael*, Springer, Berlin, 2002.

[35] P.J. Davis and P. Rabinowitz, *Methods of Numerical Integration*, 2nd ed., Academic Press, New York, 1984.

[36] L. Devroye, *Non-Uniform Random Variate Generation*, Springer, New York, 1986.

[37] L. Devroye, Binary search trees based on Weyl and Lehmer sequences, *Monte Carlo and Quasi-Monte Carlo Methods 1996* (H. Niederreiter, P. Hellekalek, G. Larcher, and P. Zinterhof, eds.), Lecture Notes in Statistics, Vol. **127**, pp. 40–65, Springer, New York, 1998.

[38] J. Dick and H. Niederreiter, On the exact t-value of Niederreiter and Sobol' sequences, *J. Complexity* **24**, 572–581 (2008).

[39] J. Dick and F. Pillichshammer, *Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration*, Cambridge University Press, Cambridge, 2010.

[40] W. Diffie and M.E. Hellman, New directions in cryptography, *IEEE Trans. Inform. Theory* **22**, 644–654 (1976).

[41] M. Drmota and R.F. Tichy, *Sequences, Discrepancies and Applications*, Lecture Notes in Mathematics, Vol. **1651**, Springer, Berlin, 1997.

[42] R. Eckhardt, Stan Ulam, John von Neumann, and the Monte Carlo method, *From Cardinals to Chaos: Reflections on the Life and Legacy of Stanislaw Ulam* (N.G. Cooper, ed.), pp. 131–137, Cambridge University Press, Cambridge, 1989.

[43] J. Eichenauer and J. Lehn, A non-linear congruential pseudo random number generator, *Statistical Papers* **27**, 315–326 (1986).

[44] J. Eichenauer-Herrmann, Inversive congruential pseudorandom numbers avoid the planes, *Math. Comp.* **56**, 297–301 (1991).

[45] J. Eichenauer-Herrmann, Statistical independence of a new class of inversive congruential pseudorandom numbers, *Math. Comp.* **60**, 375–384 (1993).

[46] J. Eichenauer-Herrmann and H. Niederreiter, Digital inversive pseudorandom numbers, *ACM Trans. Modeling and Computer Simulation* **4**, 339–349 (1994).

[47] A. Enge, *Elliptic Curves and Their Applications to Cryptography: An Introduction*, Kluwer Academic Publishers, Boston, 1999.

[48] A.B. Evans, The existence of strong complete mappings of finite groups: a survey, *Discrete Math.* **313**, 1191–1196 (2013).

[49] H. Faure, Discrépance de suites associées à un système de numération (en dimension s), *Acta Arith.* **41**, 337–351 (1982). (French)

[50] H. Faure and P. Kritzer, New star discrepancy bounds for (t, m, s)-nets and (t, s)-sequences, *Monatsh. Math.* **172**, 55–75 (2013).

[51] G.S. Fishman, *Monte Carlo: Concepts, Algorithms, and Applications*, Springer, New York, 1996.

[52] M.Z. Garaev, The sum-product estimate for large subsets of prime fields, *Proc. Amer. Math. Soc.* **136**, 2735–2739 (2008).

[53] A. Garcia and H. Stichtenoth (eds.), *Topics in Geometry, Coding Theory and Cryptography*, Springer, Dordrecht, 2007.

[54] J.E. Gentle, *Random Number Generation and Monte Carlo Methods*, 2nd ed., Springer, New York, 2003.

[55] P. Glasserman, *Monte Carlo Methods in Financial Engineering*, Springer, New York, 2004.

[56] A. Glibichuk and M. Rudnev, On additive properties of product sets in an arbitrary finite field, *J. Analyse Math.* **108**, 159–170 (2009).

[57] M.J.E. Golay, Notes on digital coding, *Proc. Inst. Radio Engineers* **37**, 657 (1949).

[58] J.H. Halton, On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals, *Numer. Math.* **2**, 84–90 (1960); Erratum, *ibid.* **2**, 196 (1960).

[59] J.M. Hammersley, Monte Carlo methods for solving multivariable problems, *Ann. New York Acad. Sci.* **86**, 844–874 (1960).

[60] J.M. Hammersley and D.C. Handscomb, *Monte Carlo Methods*, Methuen, London, 1964.

[61] R.W. Hamming, Error detecting and error correcting codes, *Bell System Tech. Journal* **29**, 147–160 (1950).

[62] G.H. Hardy and E.M. Wright, *An Introduction to the Theory of Numbers*, 6th ed., Clarendon Press, Oxford, 2008.

[63] E. Hlawka, Funktionen von beschränkter Variation in der Theorie der Gleichverteilung, *Ann. Mat. Pura Appl.* (IV) **54**, 325–333 (1961). (German)

[64] E. Hlawka, Zur angenäherten Berechnung mehrfacher Integrale, *Monatsh. Math.* **66**, 140–151 (1962). (German)

[65] A. Hocquenghem, Codes correcteurs d'erreurs, *Chiffres* **2**, 147–156 (1959). (French)

[66] R. Hofer and H. Niederreiter, A construction of (t, s)-sequences with finite-row generating matrices using global function fields, *Finite Fields Appl.* **21**, 97–110 (2013).

[67] R. Hofer and H. Niederreiter, Vandermonde nets, *Acta Arith.* **163**, 145–160 (2014).

[68] C. Hooley, On Artin's conjecture, *J. Reine Angew. Math.* **225**, 209–220 (1967).

[69] K.J. Horadam, *Hadamard Matrices and Their Applications*, Princeton University Press, Princeton, NJ, 2007.

[70] K.J. Horadam, Hadamard matrices and their applications: progress 2007–2010, *Cryptography and Communications* **2**, 129–154 (2010).

[71] W. Hörmann, J. Leydold, and G. Derflinger, *Automatic Nonuniform Random Variate Generation*, Springer, Berlin, 2004.

[72] T. Itoh and S. Tsujii, A fast algorithm for computing multiplicative inverses in GF(2^m) using normal bases, *Information and Computation* **78**, 171–177 (1988).

[73] F. James, A review of pseudorandom number generators, *Computer Physics Comm.* **60**, 329–344 (1990).

[74] D. Jungnickel, *Finite Fields: Structure and Arithmetics*, Bibliographisches Institut, Mannheim, 1993.

[75] D. Kahn, *The Codebreakers*, Macmillan Publishing Company, New York, 1967.

[76] M.H. Kalos and P.A. Whitlock, *Monte Carlo Methods*, 2nd ed., Wiley, New York, 2008.

[77] A.Y. Khinchin, *Three Pearls of Number Theory*, Graylock Press, Rochester, NY, 1952.

[78] A. Kholosha and A. Pott, Bent and related functions, *Handbook of Finite Fields* (G.L. Mullen and D. Panario, eds.), pp. 262–273, CRC Press, Boca Raton, FL, 2013.

[79] J. Kiefer, On large deviations of the empiric d.f. of vector chance variables and a law of the iterated logarithm, *Pacific J. Math.* **11**, 649–660 (1961).

[80] A. Klappenecker and M. Rötteler, Constructions of mutually unbiased bases, *Finite Fields and Applications* (G.L. Mullen, A. Poli, and H. Stichtenoth, eds.), Lecture Notes in Computer Science, Vol. **2948**, pp. 137–144, Springer, Berlin, 2004.

[81] D.E. Knuth, *The Art of Computer Programming*, Vol. 2: *Seminumerical Algorithms*, 3rd ed., Addison-Wesley, Reading, MA, 1998.

[82] N. Koblitz, *A Course in Number Theory and Cryptography*, 2nd ed., Graduate Texts in Mathematics, Vol. **114**, Springer, New York, 1994.

[83] N. Koblitz, *Algebraic Aspects of Cryptography*, Springer, Berlin, 1998.

[84] J.F. Koksma, Een algemeene stelling uit de theorie der gelijkmatige verdeeling modulo 1, *Mathematica B* (*Zutphen*) **11**, 7–11 (1942/43). (Dutch)

[85] N.M. Korobov, The approximate computation of multiple integrals, *Dokl. Akad. Nauk SSSR* **124**, 1207–1210 (1959). (Russian)

[86] N.M. Korobov, Properties and calculation of optimal coefficients, *Dokl. Akad. Nauk SSSR* **132**, 1009–1012 (1960). (Russian)

[87] N.M. Korobov, *Number-Theoretic Methods in Approximate Analysis*, Fizmatgiz, Moscow, 1963. (Russian)

[88] A.E. Kossovsky, *Benford's Law*, World Scientific, Singapore, 2014.

[89] P. Kritzer, Improved upper bounds on the star discrepancy of (t, m, s)-nets and (t, s)-sequences, *J. Complexity* **22**, 336–347 (2006).

[90] J.M. Kubina and M.C. Wunderlich, Extending Waring's conjecture to $471, 600, 000$, *Math. Comp.* **55**, 815–820 (1990).

[91] L. Kuipers and H. Niederreiter, *Uniform Distribution of Sequences*, Wiley, New York, 1974; reprint, Dover Publications, Mineola, NY, 2006.

[92] S. Kunoff, $N!$ has the first digit property, *Fibonacci Quarterly* **25**, 365–367 (1987).

[93] K. Lally and P. Fitzpatrick, Algebraic structure of quasicyclic codes, *Discrete Applied Math.* **111**, 157–175 (2001).

[94] C.F. Laywine and G.L. Mullen, *Discrete Mathematics Using Latin Squares*, Wiley, New York, 1998.

[95] P. L'Ecuyer and P. Hellekalek, Random number generators: selection criteria and testing, *Random and Quasi-Random Point Sets* (P. Hellekalek and G. Larcher, eds.), Lecture Notes in Statistics, Vol. **138**, pp. 223–265, Springer, New York, 1998.

[96] D.H. Lehmer, Mathematical methods in large-scale computing units, *Proc. 2nd Symposium on Large-Scale Digital Calculating Machinery* (Cambridge, MA, 1949), pp. 141–146, Harvard University Press, Cambridge, MA, 1951.

[97] C. Lemieux, *Monte Carlo and Quasi-Monte Carlo Sampling*, Springer, New York, 2009.

[98] G. Leobacher and F. Pillichshammer, *Introduction to Quasi-Monte Carlo Integration and Applications*, Birkhäuser and Springer International, Heidelberg, 2014.

[99] K.H. Leung and B. Schmidt, New restrictions on possible orders of circulant Hadamard matrices, *Designs Codes Cryptography* **64**, 143–151 (2012).

[100] S. Levy, *Crypto*, Viking Penguin, New York, 2001.

[101] M. Li and P.M.B. Vitányi, Kolmogorov complexity and its applications, *Handbook of Theoretical Computer Science*, Vol. A: *Algorithms and Complexity* (J. van Leeuwen, ed.), pp. 187–254, Elsevier, Amsterdam, 1990.

[102] R. Lidl and H. Niederreiter, *Finite Fields*, Addison-Wesley, Reading, MA, 1983; reprint, Cambridge University Press, Cambridge, 1997.

[103] R. Lidl and H. Niederreiter, *Introduction to Finite Fields and Their Applications*, revised ed., Cambridge University Press, Cambridge, 1994.

[104] S. Ling, H. Niederreiter, and P. Solé, On the algebraic structure of quasi-cyclic codes IV: repeated roots, *Designs Codes Cryptography* **38**, 337–361 (2006).

[105] S. Ling and P. Solé, On the algebraic structure of quasi-cyclic codes I: finite fields, *IEEE Trans. Inform. Theory* **47**, 2751–2760 (2001).

[106] S. Ling and C.P. Xing, *Coding Theory: A First Course*, Cambridge University Press, Cambridge, 2004.

[107] F.J. MacWilliams, *Combinatorial Problems of Elementary Group Theory*, Ph.D. Thesis, Department of Mathematics, Harvard University, 1962.

[108] F.J. MacWilliams and N.J.A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, 1977.

[109] K. Mahler, On the fractional parts of the powers of a rational number. II, *Mathematika* **4**, 122–124 (1957).

[110] G. Marsaglia, Random numbers fall mainly in the planes, *Proc. Nat. Acad. Sci. U.S.A.* **61**, 25–28 (1968).

[111] M. Matsumoto and T. Nishimura, Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator, *ACM Trans. Modeling and Computer Simulation* **8**, 3–30 (1998).

[112] A. May, Computing the RSA secret key is deterministic polynomial time equivalent to factoring, *Advances in Cryptology—CRYPTO 2004* (M. Franklin, ed.), Lecture Notes in Computer Science, Vol. **3152**, pp. 213–219, Springer, Berlin, 2004.

[113] R.J. McEliece, *The Theory of Information and Coding*, Addison-Wesley, Reading, MA, 1977.

[114] W. Meidl and A. Winterhof, Linear complexity of sequences and multisequences, *Handbook of Finite Fields* (G.L. Mullen and D. Panario, eds.), pp. 324–336, CRC Press, Boca Raton, FL, 2013.

[115] A.J. Menezes, *Elliptic Curve Public Key Cryptosystems*, Kluwer Academic Publishers, Boston, 1993.

[116] A.J. Menezes, P.C. van Oorschot, and S.A. Vanstone, *Handbook of Applied Cryptography*, CRC Press, Boca Raton, FL, 1997.

[117] G.L. Miller, Riemann's hypothesis and tests for primality, *J. Computer and System Sciences* **13**, 300–317 (1976).

[118] G.L. Mullen and D. Panario (eds.), *Handbook of Finite Fields*, CRC Press, Boca Raton, FL, 2013.

[119] T. Müller-Gronbach, E. Novak, and K. Ritter, *Monte Carlo-Algorithmen*, Springer, Berlin, 2012. (German)

[120] K. Nagasaka and J.-S. Shiue, Benford's law for linear recurrence sequences, *Tsukuba J. Math.* **11**, 341–351 (1987).

[121] M.B. Nathanson, *Additive Number Theory*, Vol. 2: *Inverse Problems and the Geometry of Sumsets*, Graduate Texts in Mathematics, Vol. **165**, Springer, New York, 1996.

[122] M.B. Nathanson (ed.), *Unusual Applications of Number Theory*, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Vol. **64**, American Mathematical Society, Providence, RI, 2004.

[123] P.M. Neumann (ed.), *The Mathematical Writings of Evariste Galois*, European Mathematical Society, Zurich, 2011.

[124] H. Niederreiter, Methods for estimating discrepancy, *Applications of Number Theory to Numerical Analysis* (S.K. Zaremba, ed.), pp. 203–236, Academic Press, New York, 1972.

[125] H. Niederreiter, On the distribution of pseudo-random numbers generated by the linear congruential method. III, *Math. Comp.* **30**, 571–597 (1976).

[126] H. Niederreiter, Existence of good lattice points in the sense of Hlawka, *Monatsh. Math.* **86**, 203–219 (1978).

[127] H. Niederreiter, Quasi-Monte Carlo methods and pseudo-random numbers, *Bull. Amer. Math. Soc.* **84**, 957–1041 (1978).

[128] H. Niederreiter, The serial test for pseudo-random numbers generated by the linear congruential method, *Numer. Math.* **46**, 51–68 (1985).

[129] H. Niederreiter, Low-discrepancy point sets, *Monatsh. Math.* **102**, 155–167 (1986).

[130] H. Niederreiter, Point sets and sequences with small discrepancy, *Monatsh. Math.* **104**, 273–337 (1987).

[131] H. Niederreiter, Low-discrepancy and low-dispersion sequences, *J. Number Theory* **30**, 51–70 (1988).

[132] H. Niederreiter, Statistical independence of nonlinear congruential pseudorandom numbers, *Monatsh. Math.* **106**, 149–159 (1988).

[133] H. Niederreiter, Low-discrepancy point sets obtained by digital constructions over finite fields, *Czechoslovak Math. J.* **42**, 143–166 (1992).

[134] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*, SIAM, Philadelphia, 1992.

[135] H. Niederreiter, Improved error bounds for lattice rules, *J. Complexity* **9**, 60–75 (1993).

[136] H. Niederreiter, New developments in uniform pseudorandom number and vector generation, *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing* (H. Niederreiter and P.J.-S. Shiue, eds.), pp. 87–120, Springer, New York, 1995.

[137] H. Niederreiter, The multiple-recursive matrix method for pseudorandom number generation, *Finite Fields Appl.* **1**, 3–30 (1995).

[138] H. Niederreiter, The independence of two randomness properties of sequences over finite fields, *J. Complexity* **28**, 154–161 (2012).

[139] H. Niederreiter and G. Pirsic, Duality for digital nets and its applications, *Acta Arith.* **97**, 173–182 (2001).

[140] H. Niederreiter and K.H. Robinson, Complete mappings of finite fields, *J. Austral. Math. Soc. Ser. A* **33**, 197–212 (1982).

[141] H. Niederreiter and I.E. Shparlinski, On the distribution and lattice structure of nonlinear congruential pseudorandom numbers, *Finite Fields Appl.* **5**, 246–253 (1999).

[142] H. Niederreiter and I.E. Shparlinski, On the distribution of inversive congruential pseudorandom numbers in parts of the period, *Math. Comp.* **70**, 1569–1574 (2001).

[143] H. Niederreiter and I.E. Shparlinski, Recent advances in the theory of nonlinear pseudorandom number generators, *Monte Carlo and Quasi-Monte Carlo Methods 2000* (K.-T. Fang, F.J. Hickernell, and H. Niederreiter, eds.), pp. 86–102, Springer, Berlin, 2002.

[144] H. Niederreiter and I.H. Sloan, Quasi-Monte Carlo methods with modified vertex weights, *Numerical Integration IV* (H. Braß and G. Hämmerlin, eds.), International Series of Numerical Mathematics, Vol. **112**, pp. 253–265, Birkhäuser, Basel, 1993.

[145] H. Niederreiter and A. Winterhof, Exponential sums for nonlinear recurring sequences, *Finite Fields Appl.* **14**, 59–64 (2008).

[146] H. Niederreiter and C.P. Xing, Low-discrepancy sequences and global function fields with many rational places, *Finite Fields Appl.* **2**, 241–273 (1996).

[147] H. Niederreiter and C.P. Xing, *Rational Points on Curves over Finite Fields: Theory and Applications*, Cambridge University Press, Cambridge, 2001.

[148] H. Niederreiter and C.P. Xing, *Algebraic Geometry in Coding Theory and Cryptography*, Princeton University Press, Princeton, NJ, 2009.

[149] H. Niederreiter and A.S.J. Yeo, Halton-type sequences from global function fields, *Science China Math.* **56**, 1467–1476 (2013).

[150] M.A. Nielsen and I.L. Chuang, *Quantum Computation and Quantum Information*, Cambridge University Press, Cambridge, 2000.

[151] I. Niven, *Irrational Numbers*, Mathematical Association of America, Washington, DC, 1956.

[152] I. Niven, H.S. Zuckerman, and H.L. Montgomery, *An Introduction to the Theory of Numbers*, 5th ed., Wiley, New York, 1991.

[153] H.J. Nussbaumer, *Fast Fourier Transform and Convolution Algorithms*, Springer, Berlin, 1981.

[154] D. Nuyens and R. Cools, Fast algorithms for component-by-component construction of rank-1 lattice rules in shift-invariant reproducing kernel Hilbert spaces, *Math. Comp.* **75**, 903–920 (2006).

[155] O. Ore, *Number Theory and Its History*, McGraw-Hill, New York, 1948.

[156] R. Overbeck and N. Sendrier, Code-based cryptography, *Post-Quantum Cryptography* (D.J. Bernstein, J. Buchmann, and E. Dahmen, eds.), pp. 95–145, Springer, Berlin, 2009.

[157] R.E.A.C. Paley, On orthogonal matrices, *J. Math. Physics* **12**, 311–320 (1933).

[158] S.H. Paskov and J.F. Traub, Faster valuation of financial derivatives, *J. Portfolio Management* **22**, no. 1, 113–120 (1995).

[159] T. Petsinis, *The French Mathematician*, Walker Publishing Co., New York, 1998.

[160] J. Pieprzyk, T. Hardjono, and J. Seberry, *Fundamentals of Computer Security*, Springer, Berlin, 2003.

[161] G. Pirsic, J. Dick, and F. Pillichshammer, Cyclic digital nets, hyperplane nets, and multivariate integration in Sobolev spaces, *SIAM J. Numer. Analysis* **44**, 385–411 (2006).

[162] V.S. Pless and W.C. Huffman (eds.), *Handbook of Coding Theory*, Elsevier, Amsterdam, 1998.

[163] M.O. Rabin, Probabilistic algorithm for testing primality, *J. Number Theory* **12**, 128–138 (1980).

[164] RAND Corporation, *One Million Random Digits and 100,000 Normal Deviates*, Free Press, Glencoe, IL, 1955.

[165] I.S. Reed and G. Solomon, Polynomial codes over certain finite fields, *J. Soc. Industrial Applied Math.* **8**, 300–304 (1960).

[166] D.L. Reiner, Enumeration in music theory, *Amer. Math. Monthly* **92**, 51–54 (1985).

[167] L. Rempe-Gillen and R. Waldecker, *Primality Testing for Beginners*, Student Mathematical Library, Vol. **70**, American Mathematical Society, Providence, RI, 2014.

[168] R.D. Richtmyer, The evaluation of definite integrals, and a quasi-Monte-Carlo method based on the properties of algebraic numbers, Report LA-1342, Los Alamos Scientific Laboratory, Los Alamos, NM, 1951.

[169] H. Riesel, *Prime Numbers and Computer Methods for Factorization*, 2nd ed., Birkhäuser, Basel, 1994.

[170] B.D. Ripley, *Stochastic Simulation*, Wiley, New York, 1987.

[171] R.L. Rivest, A. Shamir, and L. Adleman, A method for obtaining digital signatures and public-key cryptosystems, *Communications of the ACM* **21**, 120–126 (1978).

[172] A.M. Rockett and P. Szűsz, *Continued Fractions*, World Scientific, Singapore, 1992.

[173] S. Roman, *Field Theory*, Graduate Texts in Mathematics, Vol. **158**, Springer, New York, 1995.

[174] M.Yu. Rosenbloom and M.A. Tsfasman, Codes for the m-metric, *Problems Inform. Transmission* **33**, 45–52 (1997).

[175] J.W. Rosenthal, Card shuffling, *Math. Magazine* **54**, 64–67 (1981).

[176] K.F. Roth, On irregularities of distribution, *Mathematika* **1**, 73–79 (1954).

[177] W.M. Schmidt, Irregularities of distribution. VII, *Acta Arith.* **21**, 45–50 (1972).

[178] J. Schwinger, Unitary operator bases, *Proc. Nat. Acad. Sci. U.S.A.* **46**, 570–579 (1960).

[179] C.E. Shannon, A mathematical theory of communication, *Bell System Tech. Journal* **27**, 379–423, 623–656 (1948).

[180] P.W. Shor, Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer, *SIAM J. Computing* **26**, 1484–1509 (1997).

[181] I.E. Shparlinski, *Finite Fields: Theory and Computation*, Kluwer Academic Publishers, Dordrecht, 1999.

[182] I.E. Shparlinski, *Cryptographic Applications of Analytic Number Theory: Complexity Lower Bounds and Pseudorandomness*, Progress in Computer Science and Applied Logic, Vol. **22**, Birkhäuser, Basel, 2003.

[183] I.E. Shparlinski and J.F. Voloch, Visible points on curves over finite fields, *Bull. Pol. Acad. Sci. Math.* **55**, 193–199 (2007).

[184] I.E. Shparlinski and A. Winterhof, Visible points on multidimensional modular hyperbolas, *J. Number Theory* **128**, 2695–2703 (2008).

[185] Yu.A. Shreider (ed.), *The Monte Carlo Method*, Pergamon Press, Oxford, 1966.

[186] V. Sinescu and S. Joe, Good lattice rules with a composite number of points based on the product weighted star discrepancy, *Monte Carlo and Quasi-Monte Carlo Methods 2006* (A. Keller, S. Heinrich, and H. Niederreiter, eds.), pp. 645–658, Springer, Berlin, 2008.

[187] S. Singh, *The Code Book*, Doubleday, New York, 1999.

[188] R.C. Singleton, Maximum distance q-nary codes, *IEEE Trans. Inform. Theory* **10**, 116–118 (1964).

[189] I.H. Sloan and S. Joe, *Lattice Methods for Multiple Integration*, Clarendon Press, Oxford, 1994.

[190] I.M. Sobol', Distribution of points in a cube and approximate evaluation of integrals, *USSR Comp. Math. and Math. Physics* **7**, no. 4, 86–112 (1967).

[191] R.M. Solovay and V. Strassen, A fast Monte-Carlo test for primality, *SIAM J. Computing* **6**, 84–85 (1977).

[192] H. Stichtenoth, *Algebraic Function Fields and Codes*, 2nd ed., Springer, Berlin, 2009.

[193] D.R. Stinson, *Cryptography: Theory and Practice*, 3rd ed., CRC Press, Boca Raton, FL, 2006.

[194] H.C.A. van Tilborg, *Fundamentals of Cryptology*, revised ed., Kluwer Academic Publishers, Boston, 2000.

[195] A. Topuzoğlu and A. Winterhof, Pseudorandom sequences, *Topics in Geometry, Coding Theory and Cryptography* (A. Garcia and H. Stichtenoth, eds.), pp. 135–166, Springer, Dordrecht, 2007.

[196] W. Trappe and L.C. Washington, *Introduction to Cryptography with Coding Theory*, 2nd ed., Pearson Prentice Hall, Upper Saddle River, NJ, 2006.

[197] R.C. Vaughan and T.D. Wooley, Waring's problem: a survey, *Number Theory for the Millennium* (M.A. Bennett *et al.*, eds.), Vol. III, pp. 301–340, A.K. Peters, Natick, MA, 2002.

[198] S.S. Wagstaff Jr., *The Joy of Factoring*, Student Mathematical Library, Vol. **68**, American Mathematical Society, Providence, RI, 2013.

[199] L.C. Washington, *Elliptic Curves: Number Theory and Cryptography*, CRC Press, Boca Raton, FL, 2003.

[200] H. Weyl, Über die Gleichverteilung von Zahlen mod. Eins, *Math. Annalen* **77**, 313–352 (1916). (German)

[201] A. Winterhof, Linear complexity and related complexity measures, *Selected Topics in Information and Coding Theory* (I. Woungang, S. Misra, and S.C. Misra, eds.), pp. 3–40, World Scientific, Singapore, 2010.

[202] A. Winterhof, Recent results on recursive nonlinear pseudorandom number generators (invited paper), *Sequences and Their Applications—SETA 2010* (C. Carlet and A. Pott, eds.), Lecture Notes in Computer Science, Vol. **6338**, pp. 113–124, Springer, Berlin, 2010.

[203] C.P. Xing and H. Niederreiter, A construction of low-discrepancy sequences using global function fields, *Acta Arith.* **73**, 87–102 (1995).

[204] S.K. Zaremba, Good lattice points, discrepancy, and numerical integration, *Ann. Mat. Pura Appl.* (IV) **73**, 293–317 (1966).

[205] S.K. Zaremba, Some applications of multidimensional integration by parts, *Ann. Polon. Math.* **21**, 85–96 (1968).

[206] S.K. Zaremba, La méthode des “bons treillis” pour le calcul des intégrales multiples, *Applications of Number Theory to Numerical Analysis* (S.K. Zaremba, ed.), pp. 39–119, Academic Press, New York, 1972. (French)

Index