NUMBER THEORY – MATH 346 & 377
COURSE NOTES
WINTER 2021
VERSION: April 22, 2021

EYAL Z. GOREN,
MCGILL UNIVERSITY

INTRODUCTION

Continued fractions are an old subject. In a sense, to be described below, the golden-ratio, for example, is inherently a continued fraction. Also, as we shall see, the Euclidean algorithm for calculating greatest common divisors is a continued fraction algorithm in disguise. Continued fractions also appear in ancient Sanskrit manuscripts, and certain approximation to π suggest that some of the theory must have been known in ancient China, Egypt, India and Persia. Thus, in some sense, the origins of the subject are in antiquity. Mathematically, the subject perhaps started taking off with the work of L. Euler in the 18th century.

In any case, our interest in the subject is not historic.[1] Continued fractions are related to several interesting areas in number theory and outside it. They play a role in diophantine approximation as providing the best method to find good rational approximations to irrational numbers. Rational approximations to π and to square roots always played in important role, motivation ranging from studying the heavens, to constructing the pyramids. Continued fractions play a role in transcendence theory as providing many beautiful examples of transcendental numbers described by their continued fraction expansions. They provide solutions to what may perhaps be considered as the simplest non-trivial diophantine equations, $x^2 - dy^2 = 1$, the Pell equation. Instances of these equations appear in ancient time as well, for example, a problem known as the Archimedes cattle problem translates into a Pell equation with solutions so large, no one in that era was able to find, or even write down.[2]

Much later, perhaps in the mid-20th century, a new aspect of studying continued fractions developed; the study of their statistical properties: what is the measure of the set of continued fractions enjoying a given property? How do continued fractions behave on average relative to some property? And so on. At the same time, sets defined by continued fractions are examples both of fractals and of attractors of dynamical systems and so they became interesting examples, as well as a test-ground, for questions in ergodic theory and dynamical systems.

We will touch on all these subjects. Prove many results, cite some others, and hopefully create enough interest so that the reader will choose to pursue the theory further.

---

[1] Those interested in learning more about the history of continued fractions can consult the book by C. Brezinski, *History of Continued Fractions and Padé Approximants*, although his discussion ends at the beginning of WWII, and so almost none of the modern, and exciting, developments in this area is presented. Nonetheless, even a quick browse of the book will convince the reader of the immense scope and applications of continued fractions.

[2] If you are interested in learning more about this, nothing beats the article of H. Lenstra, "Solving the Pell Equation", *Notices AMS*, 49 (2): 182–192.

CONTINUED FRACTIONS

References. The topic of continued fractions is covered in most books on number theory. Two good references are:

(1) A. Ya. Khinchin. *Continued Fractions*. Dover Publications; 1st edition (May 14 1997).
(2) Godfrey H. Hardy, Edward M. Wright. *An Introduction to the Theory of Numbers*. Oxford University Press; Sixth edition (July 31 2008).

That said, there is material presented in these notes that is not taken from these text books and the text books contain material not in these notes. In principle, the notes should suffice for the course. There is no need to purchase these text books, but if you do want to consult additional resources, you can start with those. For the part concerning Hausdorff dimension I am following mainly the book

(3) Kenneth Falconer: *Fractal Geometry: Mathematical Foundations and Applications*. Wiley; 3rd edition (Dec 31 2013)

1. INTRODUCTION

The constant π is not a rational number; $\pi \neq \frac{a}{b}$ for any $a, b \in \mathbb{Q}$. In fact, π is not even an algebraic number: for any non-zero polynomial $f(x) \in \mathbb{Q}[x], f(x) = a_n x^n + \cdots + a_1 x + a_0$, we have $f(\pi) \neq 0$. Explicitly, for any rational numbers a_i, not all zero, and a non-negative integer n,

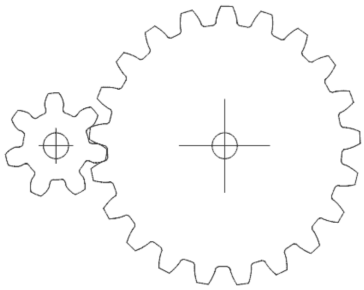
$$a_n \pi^n + \cdots + a_1 \pi + a_0 \neq 0.$$

On the other hand, we can find excellent rational approximations to π:

$$\pi = 3.1415926\ldots, \quad \frac{22}{7} = 3.1428\ldots, \quad \frac{355}{113} = 3.1415929\ldots.$$

It is notable that relative to the size of the denominator, these are excellent approximations. Continued fractions provide a method to find all such optimal approximations.

We will see that the theory of continued fractions is rich and full of mathematical ideas and applications. It also has applications to mechanics. For example, for designing watches. Suppose we want to create a hand on a watch that completes a sweep of it every π-seconds. If we take two cogs, one with 7 teeth and the other with 22. Then by rotating the 7-teeth cog at a rate of one revolution per second, the other cog will complete a revolution in 22/7 seconds, which is very nearly π.



1.1. **Notation.** Given $a_0 \in \mathbb{Z}$ and $a_1, \dots, a_N \in \mathbb{N}^+ := \{1, 2, 3, \dots\}$ we let

$$[a_0, a_1, a_2, \dots, a_N] = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \ddots + \cfrac{1}{a_{N-1} + \cfrac{1}{a_N}}}}}$$

We call such an expression a **finite continued fraction**. The a_i are called **partial quotients**. For example,

$$[3] = 3, \qquad [3,7] = 3 + \frac{1}{7} = \frac{22}{7}, \qquad [3,7,15] = 3 + \cfrac{1}{7 + \cfrac{1}{15}} = \frac{333}{106},$$

and

$$[3,7,15,1] = 3 + \cfrac{1}{7 + \cfrac{1}{15 + \cfrac{1}{1}}} = \frac{355}{113}.$$

The notation for continuous fractions is "expensive"; it takes a lot of room on the page. We therefore introduce a more compact notation. We will also use the notation

$$a_0 + \frac{1}{a_1+}\rfloor\frac{1}{a_2+}\rfloor \dots \frac{1}{a_{N-1}+}\rfloor\frac{1}{a_N} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \ddots + \cfrac{1}{a_{N-1} + \frac{1}{a_N}}}}}$$

1.2. **Sample of results.** Here is a sample of some of the theorems we will prove in this part of the course.

Theorem 1.2.1. *Let $a_0 \in \mathbb{Z}$, $a_1, a_2, a_3, \dots \in \mathbb{N}^+$. The series*

$$[a_0], [a_0, a_1], [a_0, a_1, a_2], \dots$$

converges to an irrational real number θ, which we denote

$$[a_0, a_1, a_2, \dots].$$

We call such an expression an **infinite continued fraction**. *The rational number $[a_0, a_1, \dots, a_k]$ is called the k-th* **convergent**.

Conversely, given $\theta \in \mathbb{R} - \mathbb{Q}$, there are unique $a_0 \in \mathbb{Z}, a_i \in \mathbb{N}^+$, such that

$$\theta = [a_0, a_1, a_2, \dots],$$

namely, such that the sequence $[a_0], [a_0, a_1], [a_0, a_1, a_2], \dots$ converges to θ.

Remark 1.2.2. Continued fractions are a tool useful for very particular purposes (and we will see some of them later). For many common purposes, such as recognizing whether a particular number θ is algebraic, they are not useful at all. In general, it is impossible to tell whether θ, given by a continuous fraction satisfies, for example, a cubic polynomial with rational coefficients. That said, there are some notable exceptions: (1) θ is rational if and only if it has a

finite continued fraction expansion. (2) θ is quadratic over $\mathbb{Q}$ if and only if its continued fraction expression is eventually periodic. Namely, of the form

$$[a_0, \dots, a_n, b_1, \dots, b_m, b_1 \dots, b_m, b_1, \dots, b_m, \dots].$$

(3) Also, there are many theorems that guarantee that continued fractions of a particular sort correspond to transcendental numbers; that is, to numbers that are not algebraic. This is a subject of on-going research, but a sample result in this area is the following: Let $m \geq 1$. Let b_0 be an integer, b_i, c_i, d_i be positive integers such that at least one of $d_1, \dots, d_m$ is not zero, then

$$\alpha = [b_0; b_1, \dots, b_s, \overline{(c_1 + \lambda d_1, \dots, c_m + \lambda d_m)}_{\lambda=0}^{\infty}]$$

is a transcendental number. Here the notation is for blocks of natural numbers $c_1, \dots, c_m, c_1 + d_1, \dots, c_m + d_m, c_1 + 2d_1, \dots, c_m + 2d_m, c_1 + 3d_1, \dots, c_m + 3d_m, \dots$.

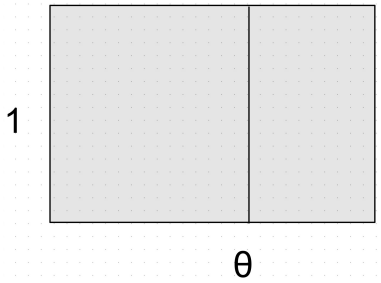
Example 1.2.3. What is $\theta = [1, 1, 1, \dots]$? We have

$$\theta = 1 + \frac{1}{1+}\rfloor\frac{1}{1+}\rfloor \dots \frac{1}{1+}\rfloor\frac{1}{+\dots} = 1 + \frac{1}{\theta}.$$

Or, simpler, $[1, 1, 1, \dots] = [1, [1, 1, 1, \dots]]$. It follows that $\theta^2 - \theta - 1 = 0$ and so

$$\theta = \frac{1 + \sqrt{5}}{2},$$

is the **golden ratio**. If you take a square of side of length 1 and extend it to a rectangle R whose shorter side is 1 and its longer side is θ, such that the rectangle formed as a difference between R and the square has the same proportion you find that $\theta : 1 = 1 : (\theta - 1)$ and so θ satisfies $\theta^2 - \theta - 1 = 0$. This is the classic definition of the golden ratio. The golden ratio is found throughout nature, architecture, art and music - from the structure of a sunflower to the ancient Parthenon in Greece; it is a proportion that is suitable for repeated and balanced structures.



Example 1.2.4. Let $\theta = \sqrt[3]{5}$. Then, $\theta = [1, 1, 2, 2, 4, 3, 3, \dots]$ and it seems that perhaps some sort of pattern will emerge. However, if we continue the development we find that

$$\sqrt[3]{5} = [1, 1, 2, 2, 4, 3, 3, 1, 5, 1, 1, 4, 10, 17, 1, 14, 1, 1, 3052, 1, 1, 1, 1, 1, 1, 2, 2, 1, 3, 2, \dots],$$

and the "random" appearance of 3052 essentially squashes all hope.

Exercise 1.2.5. What is the continued fraction $[1, 2, 3, 1, 2, 3, 1, 2, 3, \dots]$?

Exercise 1.2.6. Let a be a positive integer. What is the continued fraction expansion of the positive root of $x^2 - ax - 1$?

Exercise 1.2.7. Let a be a positive integer. What is the continued fraction expansion of the positive root of $x^2 + ax - 1$?

Especially with the last questions you may want to experiment a bit before forming a guess that you should then proceed to prove. There are many on-line continued fractions calculators and almost any mathematical software, such as Mathematica, Maple, Matlab, has such a package for continued fractions. You should be careful, especially with on-line calculators, that

they are precise enough. A free software, that is very good for number theory, is the PARI-GP software that is available for free from pari.math.u-bordeaux.fr

In PARI, the command

```
y = contfrac(Pi, 20)
```

will return the first 20 partial quotients $[a_0, \ldots, a_{19}]$ of the continued fraction expression for π.

$$[3, 7, 15, 1, 292, 1, 1, 1, 2, 1, 3, 1, 14, 2, 1, 1, 2, 2, 2, 2]$$

On the other hand, `y = contfrac(Pi, 100)` will return

$$[3, 7, 15, 1, 292, 1, 1, 1, 2, 1, 3, 1, 14, 2, 1, 1, 2, 2, 2, 2, 1, 84, 2, 1, 1, 15, 3, 13, 1, 4, 2, 6, 6]$$

which doesn't have 100 partial quotients, because Pari knows that it loses precision. If you increase the precision using

```
\p 200
```

and run `y = contfrac(Pi, 100)` again, you will get a continued fraction for π with 100 partial quotients. Running then the command

```
contfracpnqn(y, 3)
```

will return the convergents $p_0/q_0, \ldots, p_3/q_3$ (to be discussed below).

$$[3 \quad 22 \quad 333 \quad 355]$$
$$[1 \quad 7 \quad 106 \quad 113]$$

For a real number r introduce the **integer part function**

$$\lfloor r \rfloor = \max\{n \in \mathbb{Z} : n \leq r\}.$$

The **fractional part** $\{r\}$ is then defined by the identity

$$r = \lfloor r \rfloor + \{r\}.$$

In particular, $0 \leq \{r\} < 1$.

Theorem 1.2.8. *The continued fraction of $\theta \in \mathbb{R} \setminus \mathbb{Q}$ is obtained as follows.*

$$\theta = \lfloor \theta \rfloor + \{\theta\} = a_0 + \frac{1}{1/\{\theta\}} = a_0 + \frac{1}{a_1'} = a_0 + \frac{1}{a_1 + \{a_1'\}},$$

where $a_1 = \lfloor a_1' \rfloor$. Let $a_2' = 1/\{a_1'\}$ and $a_2 = \lfloor a_2' \rfloor$; we find that

$$\theta = a_0 + \frac{1}{a_1 + \frac{1}{a_2'}} = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \{a_2'\}}} = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{1/\{a_2'\}}}} == a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3'}}}$$

That is, if

$$\theta = [a_0, a_1, \ldots, a_N, a_{N+1}'],$$

where $a_0 \in \mathbb{Z}, a_1, \ldots, a_N \in \mathbb{N}^+, 1 < a_{N+1}' \in \mathbb{R}$, then write

$$a_{N+1} = \lfloor a_{N+1}' \rfloor, \quad a_{N+2}' = 1/\{a_{N+1}'\},$$

to find that

$$\theta = [a_0, a_1, \ldots, a_{N+1}, a_{N+2}'].$$

The convergents $[a_0], [a_0, a_1], [a_0, a_1, a_2], \ldots$ satisfy

$$\lim_{k \to \infty} [a_0, a_1, \ldots, a_k] = \theta.$$

Example 1.2.9. We have $\sqrt{8} = 2.828427\ldots$ and we find

$$\sqrt{8} = 2 + 0.828427\cdots = 2 + \frac{1}{1.207106\ldots} = 1 + \frac{1}{1 + \frac{1}{4.828422\ldots}}$$
$$= 1 + \frac{1}{1 + \frac{1}{4 + \frac{1}{1.207112\ldots}}} = \cdots = [2, 1, 4, 1, \ldots].$$

Example 1.2.10. Even when θ is rational, its continued fraction expansion is useful. The continued fraction expansion is defined using the same method, only that this time it ends after finitely many steps. Consider for example the reduced fraction

$$\theta = \frac{6808967}{4767456} = 1.428218\ldots.$$

This number is rather small but its expression as a fraction requires many digits (if you are not impressed, we could have easily provided a similar expression with 1000 digits in the denominator). We have

$$\theta = [1, 2, 2, 1, 57, 21, 4, 1, 5, 19].$$

The convergent $[1, 2, 2, 1, 57] = \frac{577}{404}$ is an excellent approximation:

$$\frac{577}{404} = 1.4282178\ldots$$

Of course, the point that we keep implicitly making is that continued fractions provide excellent approximations. This is a topic we will study quite closely. For example, we will prove

Theorem 1.2.11. *Let $\theta = [a_0, a_1, a_2, \ldots]$. Denote that n-th convergent as a fraction*

$$\frac{p_n}{q_n} = [a_0, a_1, a_2, \ldots, a_n],$$

with $(p_n, q_n) = 1, q_n > 0$. Then

$$\left|\theta - \frac{p_n}{q_n}\right| < \frac{1}{q_n^2}.$$

The series $q_0, q_1, q_2, q_3, \ldots$ is strictly monotone increasing for $n \geq 1$. For every n, either

$$\left|\theta - \frac{p_n}{q_n}\right| < \frac{1}{2q_n^2}, \quad \text{or} \quad \left|\theta - \frac{p_{n+1}}{q_{n+1}}\right| < \frac{1}{2q_{n+1}^2}.$$

We will call these optimal approximations. Furthermore, any optimal approximation to θ is a convergent of its continued fraction.

Exercise 1.2.12. Fill in the following table for $e = \exp(1)$. You may use a calculator, or a computer software. (For $1/q_n^2$ write an approximate decimal expansion.)

n	$[a_0, a_1, \ldots, a_n]$	p_n/q_n	$e - p_n/q_n$	$1/q_n^2$	optimal?
0	[2]	2/1	0.7182818...	1.0	no
1	[2, 1]	3/1	-0.2817181...	1.0	yes
2					
3					
4					
5					

Exercise 1.2.13. Using GP-PARI, or any other mathematical software, or even an online calculator (but make sure it's precise enough otherwise you may be lead to a wrong conjecture), find the continued fractions expansions of

$$\frac{e-1}{2}, \quad \frac{e^{1/2}-1}{2}, \quad \frac{e^{1/3}-1}{2}, \quad \frac{e^{1/4}-1}{2}, \ldots$$

Formulate a conjecture. (The pattern is hard to miss, and is a known theorem, in fact.)

2. Systematic development of the theory of continued fractions

2.1. **Convergence of continued fractions.** Recall our notation

$$[a_0, \ldots, a_N] = a_0 + \frac{1}{a_1+}\rfloor\frac{1}{a_2+}\rfloor \ldots \frac{1}{a_{N-1}+}\rfloor\frac{1}{a_N} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \ddots + \cfrac{1}{a_{N-1} + \frac{1}{a_N}}}}}$$

We now allow such expressions for any

$$a_0 \in \mathbb{R}, \quad a_i \in \mathbb{R}_{\geq 1}.$$

This is allowed just for the purpose of the proofs and once we have proven some of the basic theorems we will always assume that $a_0 \in \mathbb{Z}, a_i \in \mathbb{N}^+$ for $i \geq 1$.

Define real numbers p_n, q_n for $n \geq 0$ by the formulas

$$\begin{array}{llllll} p_0 = a_0 & p_1 = a_1a_0 + 1 & p_2 = a_2a_1a_0 + a_2 + a_0 & \ldots & p_n = a_np_{n-1} + p_{n-2} & \ldots \\ q_0 = 1 & q_1 = a_1 & q_2 = a_2a_1 + 1 & \ldots & q_n = a_nq_{n-1} + q_{n-2} & \ldots \end{array} \tag{1}$$

The recursive formula for p_n, q_n already holds for $n = 2$, in fact. The number p_n/q_n is called the **n-th convergent**.

Lemma 2.1.1. *For every n,*

$$[a_0, \ldots, a_n] = \frac{p_n}{q_n}.$$

Proof. We prove that by induction on n. For $n = 0$, we have

$$[a_0] = a_0 = \frac{p_0}{q_0}.$$

For $n = 1$, we have

$$[a_0, a_1] = a_0 + \frac{1}{a_1} = \frac{a_1 a_0 + 1}{a_1} = \frac{p_1}{q_1}.$$

For $n = 2$, we have

$$[a_0, a_1, a_2] = a_0 + \frac{1}{a_1 + \rfloor} \frac{1}{a_2} = a_0 + \frac{a_2}{a_1 a_2 + 1} = \frac{a_0 a_1 a_2 + a_2 + a_0}{a_1 a_2 + 1} = \frac{p_2}{q_2}.$$

Let us assume the result now for all $n \leq m$ for some $m \geq 2$ and prove it for $m + 1$. We have,

$$[a_0, \dots, a_m, a_{m+1}] = [a_0, \dots, a_m + \frac{1}{a_{m+1}}].$$

Let us denote the k-th convergent of $[a_0, \dots, a_m + \frac{1}{a_{m+1}}]$ by $\frac{\tilde{p}_k}{\tilde{q}_k}$. Note that $\frac{\tilde{p}_k}{\tilde{q}_k}$ depends only on the $k + 1$ first partial quotients (i.e. only on $[a_0, \dots, a_k]$ if $k < m$), and so, $\frac{\tilde{p}_k}{\tilde{q}_k} = \frac{p_k}{q_k}$ for $k < m$. At any rate, by induction,

$$\begin{aligned}
[a_0, \dots, a_m + \frac{1}{a_{m+1}}] &= \frac{\tilde{p}_m}{\tilde{q}_m} \\
&= \frac{(a_m + \frac{1}{a_{m+1}})\tilde{p}_{m-1} + \tilde{p}_{m-2}}{(a_m + \frac{1}{a_{m+1}})\tilde{q}_{m-1} + \tilde{q}_{m-2}} \\
&= \frac{(a_m + \frac{1}{a_{m+1}})p_{m-1} + p_{m-2}}{(a_m + \frac{1}{a_{m+1}})q_{m-1} + q_{m-2}} \\
&= \frac{a_{m+1}(a_m p_{m-1} + p_{m-2}) + p_{m-1}}{a_{m+1}(a_m q_{m-1} + q_{m-2}) + q_{m-1}} \\
&= \frac{a_{m+1} p_m + p_{m-1}}{a_{m+1} q_m + q_{m-1}} \\
&= \frac{p_{m+1}}{q_{m+1}}.
\end{aligned}$$

□

Example 2.1.2. Recall the Fibonacci numbers,

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 55, \dots$$

defined recursively by $a_0 = a_1 = 1$ and $a_{n+2} = a_{n+1} + a_n$.

Consider

$$\theta = [1, 1, 1, 1, \dots].$$

By the lemma, taking $n + 1$ times 1, we have

$$[1, 1, \dots, 1] = \frac{p_n}{q_n}.$$

By definition, we have

$p_0 = 1 \quad p_1 = 2 \quad p_2 = 3 \quad p_3 = 5 \quad \dots \quad p_n = (n+1)^{st}$–Fibonacci number

$q_0 = 1 \quad q_1 = 1 \quad q_2 = 2 \quad q_3 = 3 \quad \dots \quad q_n = n^{th}$–Fibonacci number

We will prove that $p_n/q_n \to \theta = [1,1,1,\dots]$ and, as we have seen, $\theta = \frac{1+\sqrt{5}}{2}$ is the golden ratio. It follows that the ratio between consecutive Fibonacci numbers converges to the golden ratio and, in fact, quite rapidly. Already,

$$\left|\frac{1+\sqrt{5}}{2} - \frac{34}{21}\right| < \frac{1}{21^2} \approx 0.002.$$

In fact, $\left|\frac{1+\sqrt{5}}{2} - \frac{34}{21}\right| = 0.00101\dots$ (34/21 is an optimal approximation).

Lemma 2.1.3. *We have*

$$p_n q_{n-1} - p_{n-1} q_n = (-1)^{n-1}, \quad n \geq 1.$$

Proof. We prove this by induction. For $n = 1$, we have

$$p_1 q_0 - p_0 q_1 = (a_1 a_0 + 1) - a_0 a_1 = 1 = (-1)^{1-1}.$$

For $n > 1$, we have

$$\begin{aligned} p_n q_{n-1} - p_{n-1} q_n &= (a_n p_{n-1} + p_{n-2}) q_{n-1} - p_{n-1}(a_n q_{n-1} + q_{n-2}) \\ &= -(p_{n-1} q_{n-2} - q_{n-1} p_{n-2}) = -(-1)^{n-2} = (-1)^{n-1}. \end{aligned}$$

□

Corollary 2.1.4. *Assume that all the a_i are integers, then so are p_n, q_n for all n, and*

$$\gcd(p_n, q_n) = 1.$$

Proof. The fact that p_n and q_n are integers is clear from the recursive formulas in (1). It follows from Lemma 2.1.3 that any common divisor of p_n and q_n divides 1, hence $(p_n, q_n) = 1$. □

By dividing by $q_n q_{n-1}$ in Lemma 2.1.3, we conclude the following.

Corollary 2.1.5. *We have for all $n \geq 1$,*

$$\left|\frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}}\right| = \frac{1}{q_n q_{n-1}}.$$

The importance of this corollary is that it allows us to show that continued fractions converge. More precisely:

Corollary 2.1.6. *Suppose that $a_0 \in \mathbb{Z}$, $a_i \in \mathbb{N}^+$. The series $\{q_n\}_{n=0}^{\infty}$ is a series of positive integers that is strictly monotone increasing for $n \geq 1$; the series of convergents*

$$\frac{p_0}{q_0}, \frac{p_1}{q_1}, \frac{p_2}{q_2}, \dots$$

is a Cauchy series, hence converges to some $\theta \in \mathbb{R}$ that we denote

$$[a_0, a_1, a_2, \dots].$$

Proof. We have $q_0 = 1, q_1 = a_1 \geq 1$ and $q_n = a_{n-1} q_{n-1} + q_{n-2}$. It follows that q_n is a strictly increasing sequence of integers for $n \geq 1$ and hence $q_n \geq n$ for $n \geq 1$. (A better estimate is provided by exercise 2.1.7 below.)

For $N \geq n \geq 1$, we have

$$\left|\frac{p_N}{q_N} - \frac{p_n}{q_n}\right| \leq \sum_{k=n}^{N-1} \left|\frac{p_{k+1}}{q_{k+1}} - \frac{p_k}{q_k}\right| = \sum_{k=n}^{N-1} \frac{1}{q_{k+1} q_k} \leq \sum_{k=n}^{N-1} \frac{1}{q_k^2} \leq \sum_{k=n}^{N-1} \frac{1}{k^2}.$$

Since $\sum_{k=1}^{\infty} \frac{1}{k^2}$ converges (to $\frac{\pi^2}{6}$, in fact), it follows that the series $\{p_n/q_n\}$ is Cauchy. □

Exercise 2.1.7. Let $[a_0, a_1, a_2, \dots]$ be a continued fraction, where $a_0 \in \mathbb{Z}, a_i \in \mathbb{N}^+, i = 1, 2, 3, \dots$. Prove that

$$q_n \geq 2^{\frac{n-1}{2}}.$$

The following lemma will allow us a very good understanding as to *how* does the series $\{p_n/q_n\}$ converge.

Lemma 2.1.8. *For all $n \geq 2$ we have*

$$p_n q_{n-2} - p_{n-2} q_n = (-1)^n a_n.$$

Proof. We have

$$\frac{p_n}{q_n} - \frac{p_{n-2}}{q_{n-2}} = (\frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}}) + (\frac{p_{n-1}}{q_{n-1}} - \frac{p_{n-2}}{q_{n-2}}) = \frac{(-1)^{n-1}}{q_n q_{n-1}} + \frac{(-1)^{n-2}}{q_{n-1} q_{n-2}} = (-1)^n \cdot \frac{q_n - q_{n-2}}{q_n q_{n-1} q_{n-2}}.$$

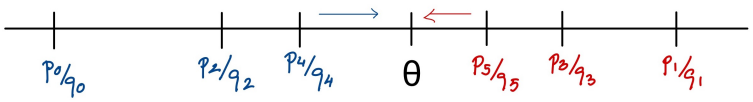
From the recursive formulas, we have $q_n - q_{n-2} = a_n q_{n-1}$. Substituting in the equation above we find

$$\frac{p_n}{q_n} - \frac{p_{n-2}}{q_{n-2}} = \frac{(-1)^n a_n}{q_n q_{n-2}}.$$

□

Corollary 2.1.9. *Suppose that $a_i \in \mathbb{N}^+$ for all $i \geq 1$ and let $\theta = [a_0, a_1, a_2, \dots]$ then we have*

$$\frac{p_0}{q_0} < \frac{p_2}{q_2} < \cdots < \theta < \cdots < \frac{p_3}{q_3} < \frac{p_1}{q_1}.$$



Proof. Given that we know:

- $p_n/q_n \to \theta$,
- $p_n/q_n - p_{n-2}q_{n-2}$ has the same sign as $(-1)^n$,
- $\left|p_n/q_n - p_{n-2}/q_{n-2}\right| > 0$,

this is the only possibility. □

Corollary 2.1.10. *For every n we have*

$$\left|\theta - \frac{p_n}{q_n}\right| < \frac{1}{q_n q_{n+1}}.$$

Proof. By Corollary 2.1.5,

$$\left|\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n}\right| = \frac{1}{q_n q_{n+1}}.$$

But note that by Corollary 2.1.9, θ is strictly between $\frac{p_n}{q_n}$ and $\frac{p_{n+1}}{q_{n+1}}$. □

2.2. **Uniqueness of continued fractions.** For $a_0 \in \mathbb{Z}$ and $a_i \in \mathbb{N}^+$ we know that the expression

$$\theta = [a_0, a_1, a_2, \dots]$$

makes sense – it is the limit of the convergents $\frac{p_k}{q_k} = [a_0, a_1, a_2, \dots, a_k]$ as $k \to \infty$. We will later prove that every irrational number has such an expression as an infinite continued fraction. Let us now prove the uniqueness of this expression.

Theorem 2.2.1. *Suppose that $a_0, b_0 \in \mathbb{Z}$, $a_i, b_i \in \mathbb{N}^+$ for $i \geq 1$ and*

$$\theta = [a_0, a_1, a_2, \dots] = [b_0, b_1, b_2, \dots].$$

Then,

$$a_i = b_i, \quad \forall i \geq 0.$$

Proof. Note that for any $n \geq 0$, $[a_{n+1}, a_{n+2}, \dots]$ is a well-defined number in $\mathbb{R}_{\geq 1}$ and

$$\theta = [a_0, a_1, a_2, \dots] = [a_0, a_1, a_2, \dots, a_n, [a_{n+1}, a_{n+2}, \dots]].$$

We have then

$$[a_0, [a_1, a_2, \dots]] = [b_0, [b_1, b_2, \dots]],$$

and it is enough to show this implies $a_0 = b_0$. Indeed, as we have

$$a_0 + \frac{1}{[a_1, a_2, \dots]} = b_0 + \frac{1}{[b_1, b_2, \dots]},$$

we will be able to conclude that

$$[a_1, a_2, \dots] = [b_1, b_2, \dots],$$

and an induction argument gives $a_i = b_i$ for all i.

What we will actually prove is that if $\theta = [a_0, [a_1, a_2, \dots]]$ then a_0 is necessarily $\lfloor \theta \rfloor$. This implies $a_0 = b_0$ and we are done. Now, because

$$\theta = a_0 + \frac{1}{[a_1, a_2, \dots]},$$

the statement $a_0 = \lfloor \theta \rfloor$ is equivalent to the statement

$$1 < [a_1, a_2, \dots].$$

This is clear: if we use the notation $[c_0, c_1, c_2, \dots] = [a_1, a_2, \dots]$ then the 0-convergent of this continued fraction is a_1, which is greater or equal to 1, and Corollary 2.1.9 shows that $c_0 < [c_0, c_1, c_2, \dots]$. □

Exercise 2.2.2. Prove that if $a_0, b_0 \in \mathbb{Z}$, $a_i, b_i \in \mathbb{N}^+$ for $i \geq 1$, we cannot have

$$[a_0, \dots, a_n] = [b_0, b_1, b_2, \dots].$$

Exercise 2.2.3. Prove that every rational number θ has a finite continued fraction expansion

$$\theta = [a_0, a_1, \dots, a_N] \quad (a_0 \in \mathbb{Z}, a_i \in \mathbb{N}^+, i = 1, \dots, N).$$

Moreover, prove that this expansion is *unique*, up to

$$[a_0, a_1, \dots, a_N] = [a_0, a_1, \dots, a_N - 1, 1],$$

if $a_N > 1$.

The development into a continued fraction you are asked to prove in the last exercise is a consequence of the Euclidean algorithm. We provide an example, leaving it to you to write the general argument.

Consider $\theta = \frac{355}{133}$. Then you may check that

$$\theta = [3, 7, 16] = [3, 7, 15, 1].$$

On the other hand, the Euclidean algorithm for finding $1 = \gcd(355, 133)$ is the following:

$$\begin{aligned} 355 &= 3 \cdot 113 + 16 \\ 113 &= 7 \cdot 16 + 1 \\ 16 &= 16 \cdot 1 + 0 \end{aligned}$$

Notice the integers $3, 7, 16$ appearing in the process as well as in the expression $\theta = [3, 7, 16]$.

2.3. **Every real number has a continued fraction expansion.** Let us summarize what we already know (we consider here, as usual, continued fractions with $a_0 \in \mathbb{Z}, a_i \in \mathbb{N}^+$ for $i \geq 1$):

- Every rational number has a finite continued expansion, unique up to $[a_0, a_1, \ldots, a_N] = [a_0, a_1, \ldots, a_N - 1, 1]$, if $a_N > 1$.
- A rational number cannot have an infinite continued fraction expansion.
- Every infinite continued fraction defines an irrational number.
- An irrational number has at most one continued fraction expansion.

The missing piece is thus provided by our next theorem.

Theorem 2.3.1. *Every irrational number θ has an infinite continued fraction expansion.*

Proof. First, we claim that for any $n \geq 0$, every irrational number θ can be written as

$$\theta = [a_0, a_1, \ldots, a_n, a'_{n+1}],$$

where $a_0 \in \mathbb{Z}, a_i \in \mathbb{N}^+$ for $i \geq 1$ and $a'_{n+1} \in \mathbb{R}_{>1}$.

Indeed, define inductively,

$$\begin{aligned} a_0 &= \lfloor \theta \rfloor & a'_1 &= 1/\{\theta\} \\ a_1 &= \lfloor a'_1 \rfloor & a'_2 &= 1/\{a'_1\} \\ &\vdots & &\vdots \\ a_n &= \lfloor a'_n \rfloor & a'_{n+1} &= 1/\{a'_n\} \end{aligned}$$

Then, $\theta = [a_0, a'_1] = [a_0, a_1, a'_2]$ and, in general,

$$\theta = [a_0, a_1, \ldots, a_{n-1}, a'_n] = [a_0, a_1, \ldots, a_{n-1}, a_n + \frac{1}{a'_{n+1}}] = [a_0, a_1, \ldots, a_{n-1}, a_n, a'_{n+1}].$$

Now, given such a presentation

$$\theta = [a_0, a_1, \ldots, a_n, a'_{n+1}],$$

define p_i, q_i as before for $i \leq n$, using the recursive formulas (those only involve $a_0, \ldots, a_n$), and define

$$p'_{n+1} = a'_{n+1} p_n + p_{n-1}, \quad q'_{n+1} = a'_{n+1} q_n + q_{n-1}.$$

We claim that for $n \geq 1$,

$$\theta = \frac{p'_{n+1}}{q'_{n+1}}.$$

This is easy to check directly for $n = 1$. Assume the result for any finite continued fraction like that of length n. Then

$$\theta = [a_0, a_1, \ldots, a_n, a'_{n+1}] = [a_0, a_1, \ldots, a_n + \frac{1}{a'_{n+1}}] = \frac{\tilde{p}'_n}{\tilde{q}'_n},$$

where $\tilde{p}'_n, \tilde{q}'_n$ are the quantities constructed for the continued fraction $[a_0, a_1, \ldots, a_n + \frac{1}{a'_{n+1}}]$. Then, using that the convergents p_k/q_k for $[a_0, a_1, \ldots, a_n, a'_{n+1}]$ are the same as the convergents $\tilde{p}_k/\tilde{q}_k$ for $[a_0, a_1, \ldots, a_n + \frac{1}{a'_{n+1}}]$ for $k < n$, we find

$$\begin{aligned}
\theta &= \frac{\tilde{p}'_n}{\tilde{q}'_n} \\
&= \frac{(a_n + \frac{1}{a'_{n+1}})\tilde{p}_{n-1} + \tilde{p}_{n-2}}{(a_n + \frac{1}{a'_{n+1}})\tilde{q}_{n-1} + \tilde{q}_{n-2}} \\
&= \frac{(a_n + \frac{1}{a'_{n+1}})p_{n-1} + p_{n-2}}{(a_n + \frac{1}{a'_{n+1}})q_{n-1} + q_{n-2}} \\
&= \frac{a'_{n+1}(a_n p_{n-1} + p_{n-2}) + p_{n-1}}{a'_{n+1}(a_n q_{n-1} + q_{n-2}) + q_{n-1}} \\
&= \frac{a'_{n+1}p_n + p_{n-1}}{a'_{n+1}q_n + q_{n-1}} \\
&= \frac{p'_{n+1}}{q'_{n+1}}.
\end{aligned}$$

Next, we prove that for all n,

$$\left|\theta - \frac{p_n}{q_n}\right| = \frac{1}{q_n q'_{n+1}}.$$

Indeed,

$$\begin{aligned}
\theta - \frac{p_n}{q_n} &= \frac{p'_{n+1}q_n - p_n q'_{n+1}}{q_n q'_{n+1}} \\
&= \frac{(a'_{n+1}p_n + p_{n-1})q_n - p_n(a'_{n+1}q_n + q_{n-1})}{q_n q'_{n+1}} \\
&= \frac{p_{n-1}q_n - p_n q_{n-1}}{q_n q'_{n+1}} \\
&= \frac{(-1)^n}{q_n q'_{n+1}}.
\end{aligned}$$

We remark that from the definition of q'_{n+1} we have $q'_{n+1} \geq q_{n+1}$. Therefore, $\left|\theta - \frac{p_n}{q_n}\right| < \frac{1}{q_n q_{n+1}}$ (as expected, given Corollary 2.1.10). Therefore, as for $n \geq 1$ we have the estimate $q_n \geq n$, we may conclude that $\left|\theta - \frac{p_n}{q_n}\right| < \frac{1}{n^2}$. This implies that

$$\lim_{n \to \infty} \frac{p_n}{q_n} = \theta,$$

and completes the proof. □

Exercise 2.3.2. ★ Use the arguments appearing in Theorem 2.3.1 to prove the following.

Theorem 2.3.3. *Let θ be an irrational real number. Then, for all $n \geq 0$ we have*

$$\left|\theta - \frac{p_n}{q_n}\right| > \frac{1}{q_n(q_{n+1} + q_n)}.$$

3. Rational, algebraic and transcendental numbers

3.1. **Enumerating.** A set S is called **countable**, or **enumerable**, if there is a bijection $f\colon \mathbb{N} \to S$. The function f allows us to "count", or "enumerate", the elements of S, since we have $S = \{f(0), f(1), f(2), \dots\}$.[3] The rational numbers $\mathbb{Q}$ are infinite, but are countable. There is a (non-obvious) bijection

$$\mathbb{N} \to \mathbb{Q},$$

which is *the proof* that $\mathbb{Q}$ is countable.

We denote the cardinality of $\mathbb{N}$ by $\aleph_0$ ($\aleph$ being the first letter in the word "infinite" (ein - sof) in Hebrew). On the other hand, the cardinality of $\mathbb{R}$ is $2^{\aleph_0}$. Where $2^{\aleph_0}$ is *defined* as the cardinality of the set of all subsets of $\mathbb{N}$. As a subset S of $\mathbb{N}$ can be specified by its characteristic function, the function that receives the value 1 on x if $x \in S$ and 0 if $x \notin S$, the set of all subsets of $\mathbb{N}$ is the set $2^{\mathbb{N}}$; namely, the set of all functions from $\mathbb{N}$ to a set of 2 elements, $\{0,1\}$ if you will. Thus, cardinality of $2^{\mathbb{N}}$ is $2^{\aleph_0}$, per definition. It is not obvious that the cardinality of $\mathbb{R}$ is $2^{\aleph_0}$ – one would have to find some method to connect subsets of $\mathbb{N}$ to real numbers to prove that – but it is a theorem and we sketch the proof below.

A very important tool in showing that two sets have the same cardinality is the following:

Cantor-Bernstein Theorem: *If there exist injective maps $A \to B$ and $B \to A$ then A and B have the same cardinality $\left|A\right| = \left|B\right|$; that is, there exists a bijection between A and B.*

It is not hard to prove that $(0,1)$ and $\mathbb{R}$ have the same cardinality, either by using the Cantor-Bernstein Theorem, or simply by using the bijection $x \mapsto \tan(\pi(x - 1/2))$. Therefore, to see that the cardinality of $\mathbb{R}$ is $2^{\aleph_0}$ it is enough to show that the cardinality of $(0,1)$ is $2^{\aleph_0}$.

Every real number in $(0,1)$ has a unique binary expansion

$$0.a_0a_1a_2\dots,$$

where $a_i \in \{0,1\}$ and where we choose expansions ending with $1000000\cdots$ instead of expansions ending in $0111111\cdots$ to have unicity. Thus, the function $i \to a_i$ defines a subset of $\mathbb{N}$, which is $\{n \in \mathbb{N} : a_n = 1\}$. Sending a real number in $(0,1)$ to the subset of $\mathbb{N}$ determined by its binary expansion is an injection

$$(0,1) \hookrightarrow 2^{\mathbb{N}}.$$

Not all subsets of $\mathbb{N}$ arise this way. For example $\mathbb{N}$ does not arise this way since we excluded the number 1 and also the subsets of the form $\{k, k+1, k+2, \dots\}$ do not arise this way, because we excluded all binary developments ending in $11111\cdots$. But, with a little bit of extra work one overcomes these blemishes and concludes that $\left|\mathbb{R}\right| = \left|(0,1)\right| = 2^{\aleph_0}$.

Cantor's diagonal argument proves that

$$\aleph_0 < \left|\mathbb{R}\right|,$$

which is equivalent to saying that there is no bijection $\mathbb{N} \to \mathbb{R}$. This is usually proven in MATH 235 (in fact without using the rather subtle fact that $\left|\mathbb{R}\right| = 2^{\aleph_0}$), so we will not repeat it here. However, if one accepts that $\left|\mathbb{R}\right| = 2^{\aleph_0}$ then we can give a short proof.

Theorem 3.1.1. *For any set A we have $\left|A\right| < \left|2^A\right|$, where 2^A is the set of subsets of A.*

[3] Some authors are more fastidious and say S is infinitely countable/enumerable and allow also finite sets to be called countable, but we will not have a use for that as our excursion into cardinalities of sets will be brief.

Proof. Clearly $A \hookrightarrow 2^A$ by sending $a \in A$ to $\{a\}$, whence $\left|A\right| \leq \left|2^A\right|$. We need to show then that $\left|A\right| \neq \left|2^A\right|$; namely, that there is no bijection between A and 2^A. Suppose that we had a bijection

$$f\colon A \to 2^A.$$

Construct a subset B of A as follows:

$$B = \{a \in A : a \notin f(a)\}.$$

Since f is surjective, $B = f(b)$ for some $b \in A$. If $b \in B$ then $b \in f(b)$ and so $b \notin B$; if $b \notin B$ then $b \notin f(b)$ and so $b \in B$. Either way, we derive a contradiction. □

Thus, in a certain sense, *most real numbers are irrational*.

A complex number α is called **algebraic** if α solves some non-zero rational polynomial $f(x) \in \mathbb{Q}[x]$. By multiplying f by a suitable rational number we may assume that f is monic:

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0, \quad a_i \in \mathbb{Q}.$$

It is not hard to prove that the set of such polynomials in countable. Fixing n, it is in bijection with $\mathbb{Q}^n$, which is countable. A general lemma states that a countable union of countable sets is countable, and that completes the proof. Furthermore, as every polynomial has finitely many roots, a similar argument proves that the cardinality of all algebraic numbers is countable. We denote the set of algebraic numbers $\overline{\mathbb{Q}}$. It is in fact a subfield of $\mathbb{C}$.

Thus, in a sense, *most real numbers are transcendental*, meaning, they are not algebraic.

Exercise 3.1.2. ★ Prove that $\bar{\mathbb{Q}}$ is a field as follows:

(1) In general, if $F \subseteq L$ are fields and $\alpha \in L$ let

$$F[\alpha] = \{\sum_{i=0}^{n} a_i \alpha^i : a_i \in \mathbb{F}\}.$$

Namely, the set of all finite polynomial expression in α with coefficients from F. Prove that $F[\alpha]$ is a ring, and that it is also a vector space over F.

(2) Prove that $\alpha \in \mathbb{C}$ is algebraic over $\mathbb{Q}$ if and only if $\dim_{\mathbb{Q}}(\mathbb{Q}[\alpha]) < \infty$. If this is the case, prove that $\mathbb{Q}[\alpha]$ is a field and, in fact, $\mathbb{Q}[\alpha] \cong \mathbb{Q}[x]/(f(x))$, where $f(x)$ is the minimal polynomial of α (see §3.3 for this notion.)

(3) Let $\alpha, \beta \in \mathbb{C}$ be algebraic over $\mathbb{Q}$. Prove that $\dim_{\mathbb{Q}}(\mathbb{Q}[\alpha, \beta]) < \infty$, where $\mathbb{Q}[\alpha, \beta] = (\mathbb{Q}[\alpha])[\beta]$.

(4) Let $\alpha, \beta \in \mathbb{C}$ be algebraic over $\mathbb{Q}$. Prove that $-\alpha$, $\frac{1}{\alpha}$ (for $\alpha \neq 0$), $\alpha + \beta$ and $\alpha\beta$ all belong to $\mathbb{Q}[\alpha, \beta]$. Conclude that they are algebraic too.

3.2. **Measuring.** We can try and study the measure of subsets of $\mathbb{R}$ to get a sense of their size. One can define a measure μ on $\mathbb{R}$ that has the following properties. Let $\mathscr{B}$ be the smallest collection of subsets of $\mathbb{R}$ that contains all open intervals and is closed under countable unions, complements and countable intersections. It is called the **Borel σ-algebra** of $\mathbb{R}$. Every open set in $\mathbb{R}$, as well as any closed set in $\mathbb{R}$ belong to $\mathscr{B}$. $\mathscr{B}$ also contains the set of all rational numbers, and the set of all irrational numbers.

The **measure** on $\mathbb{R}$ we have in mind is a function

$$\mu\colon \mathscr{B} \to \mathbb{R}_{\geq 0} \cup \{\infty\},$$

such that:

(1) $\mu((a,b)) = \mu([a,b]) = b-a$, for all $a \leq b$.
(2) $\mu(\coprod_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.
(3) for any set $A \in \mathscr{B}$ we have

$$\mu(A) = \sup\{\mu(K) : K \subseteq A,\ K \text{ compact}\} = \inf\{\mu(U) : A \subseteq U,\ U \text{ open}\}.$$

This measure is called the **Lebesgue measure**.

It follows that if $A \subset \mathbb{R}$ is countable then $\mu(A) = 0$. Indeed, by definition there is a bijection $f\colon \mathbb{N} \to A$ and so $A = \coprod_{n=0}^{\infty}\{f(n)\} = \coprod_{n=0}^{\infty}[f(n), f(n)]$. Therefore, $\mu(A) = \sum_{n=0}^{\infty} \mu([f(n), f(n)])$ $= 0$. Thus,

$$\mu(\mathbb{Q}) = \mu(\overline{\mathbb{Q}} \cap \mathbb{R}) = 0.$$

Exercise 3.2.1. Prove that the measure of $[0,1] \setminus \mathbb{Q}$ is equal to 1. Note that this set contains no interval of positive length.

Exercise 3.2.2. Let $0 \leq \alpha \leq 1$. Find a set S contained in $[0,1]$ that has measure α, contains no interval of positive length, and is dense in $[0,1]$.

Thus, *whether we are counting sets, or measuring them, almost all real numbers are irrational, and in fact transcendental.* But can we explicitly find such numbers?

It is easy to give examples of irrational numbers:

$$\sqrt{2},\ \frac{1+\sqrt{5}}{2},\ \sqrt[3]{19},$$

are all irrational. The following is a criterion that will easily prove these statements:

Let $f(x) = a_n x^n + \cdots + a_0 \in \mathbb{Z}[x]$ be a polynomial with integer coefficients. Assume that $n > 0$ and $a_n \neq 0$. If $f(p/q) = 0$ then $p|a_0, q|a_n$.

Using this, it is easy to check that the polynomials $x^2 - 2, x^2 - x - 1$ and $x^3 - 19$ don't have any rational roots. Of course, one may try and prove stronger statements. Namely, to prove that a complex number α satisfies a $f(x) \in \mathbb{Z}[x]$ of degree greater than 1 which is irreducible over $\mathbb{Q}$, and hence in particular α cannot be rational. There aren't too many methods to prove such statements. One of the most fundamental ones is **Eisenstein's criterion**:

Suppose that $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0 \in \mathbb{Z}[x]$ is a non-constant polynomial. Suppose that there is a prime p such that $p \Big| a_i$ for all i, but $p^2 \nmid a_0$. Then f is irreducible over $\mathbb{Q}$.

3.3. **Algebraic numbers.** Let $\alpha \in \mathbb{C}$ be a non-zero algebraic number. Thus, there is a non-zero polynomial

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0 \in \mathbb{Q}[x],$$

such that $f(\alpha) = 0$. The minimal such n is called the **degree** of α.

Example 3.3.1. Every non-zero rational number a/b has degree 1; it solves the polynomial $x - a/b$. Conversely, every algebraic number of degree 1 is rational.

Every quadratic number $a + b\sqrt{d}$, where d is a square-free integer and a, b rational numbers, has degree 2. It solves the polynomial $x^2 - 2ax + (a^2 - b^2 d)$. The degree cannot be 1 because that would imply that $a + b\sqrt{d}$ is rational and hence that $\sqrt{d}$ is rational, which is not the case. So, for example, $\sqrt{2}$ and $\frac{1+\sqrt{5}}{2}$ have degree 2.

The degree of $\alpha = \sqrt[3]{19}$ is 3; α solves $x^3 - 19$, which is irreducible by Eisenstein's criterion. This implies that α has degree 3 by the following lemma.

Proposition 3.3.2. *Let α be an algebraic number of degree n and let*

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$$

be a non-zero polynomial with rational coefficients such that $f(\alpha) = 0$. Then:

(1) f is irreducible.

(2) If $g(x) \in \mathbb{Q}[x]$ is a polynomial such that $g(\alpha) = 0$ then $f \Big| g$.

(3) f is the unique monic polynomial with rational coefficients of degree n that α satisfies; it is called the **minimal polynomial** *of α.*

Proof. The second statement implies the first and third. Indeed, if $f_1(x)$ is also monic of degree n having α as a root then, by (2), $f \Big| f_1$ and so $f = f_1$ and (3) follows. Suppose f is reducible: $f = gh$, where g, h are rational polynomials. Then, $f(\alpha) = g(\alpha)h(\alpha) = 0$ and so, without loss of generality, $g(\alpha) = 0$. But then, by (2), $f \Big| g$ and this implies that h is a constant polynomial. Hence (1).

Let us prove (2) then. Let $d(x) = \gcd(f, g)$; for suitable rational polynomials $u(x), v(x)$ we have

$$d(x) = u(x)g(x) + v(x)f(x).$$

Substituting α for x we find that $d(\alpha) = 0$ and so d is not a constant polynomial. As $d \Big| f$ and f has the minimal degree possible for all polynomials that α satisfies, we must have that d has degree n as well, and so $d = f$ as both are monic. That means that $f \Big| g$. □

Corollary 3.3.3. *If $g(x)$ is a non-zero monic irreducible polynomial with rational coefficients such that $g(\alpha) = 0$ then g is the minimal polynomial of α and, in particular,* $\deg(\alpha) = \deg(g)$.

A fundamental result is that

$$\bar{\mathbb{Q}} := \{\alpha \in \mathbb{C} : \alpha \text{ is algebraic}\},$$

is a field; it is closed under addition, multiplication and taking inverses. The proof was given as Exercise 3.1.2.

3.4. **Transcendental numbers.** Either in the sense of cardinality, or of measure, the non-algebraic real numbers, the real transcendental numbers, are the overwhelming majority. However, the problem of exhibiting transcendental numbers, or proving that familiar constants are transcendental is very hard in general. Many of the results we provide below are for information. That said, we will be able to explicitly exhibit transcendental numbers based on Liouville's theorem.

3.4.1. *Seminal results concerning transcendence.* We have the following seminal results.

Theorem 3.4.1 (Hermite, 1873)**.** *e is transcendental.*

Theorem 3.4.2 (Lindemann, 1882)**.** *π is transcendental.*

Theorem 3.4.3 (Gelfond-Schneider, 1934)**.** *Let α, β be algebraic numbers such that $\alpha \neq 0, 1$ and $\beta \notin \mathbb{Q}$. Then any value of α^β is transcendental.*

Some explanation is required concerning the phrase "any value" in the Theorem. The point is that α^β is really defined as $e^{\beta \cdot \log(\alpha)}$. The theorem allows α and β to be complex numbers (in particular, α could be a negative real number); the function

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!},$$

is a well-defined analytic function and, in particular, converges for every $\alpha \in \mathbb{C}$. So, e^z is well-defined for every complex number z. On the other hand, there is no global definition for $\log(x)$ that gives a well-defined answer for all $x \in \mathbb{C}$. The best one can do is provide a value for $\log(x)$ that is well defined up to integer multiples of $2\pi i$. Thus, α^β could mean any one of a countable set of complex numbers, differing from each other by integer powers of $e^{\beta \cdot 2\pi i}$.

Example 3.4.4. It follows from the Gelfond-Schneider Theorem that $2^{\sqrt{2}}$ is transcendental.

Theorem 3.4.5 (Baker, 1960's)**.** *Let $\alpha_1, \dots, \alpha_n$ be non-zero algebraic numbers such that $\log(\alpha_1), \dots, \log(\alpha_n)$ are linearly independent over $\mathbb{Q}$. Then $1, \log(\alpha_1), \dots, \log(\alpha_n)$ are linearly independent over $\bar{\mathbb{Q}}$.*

In Baker's theorem, when we talk about $\log(\alpha_i)$ we mean by that any choice of a complex number γ_i such that $e^{\gamma_i} = \alpha_i$. This theorem is extremely powerful and Alan Baker got the Fields Medal for it, and related work, in 1970. For example, Baker's Theorem implies the Gelfond-Schneider theorem:

Suppose that α^β is algebraic. Note that $\log(\alpha)$ and $\beta \log(\alpha)$ are independent over $\mathbb{Q}$, because $a \log(\alpha) + b\beta \log(\alpha) = 0$ implies that $\beta = -a/b \in \mathbb{Q}$, which is a contradiction. Baker's theorem then states that $\log(\alpha)$ and $\beta \log(\alpha)$ are independent over $\bar{\mathbb{Q}}$, which is clearly false as

$$\beta \cdot \log(\alpha) + (-1) \cdot \beta \log(\alpha) = 0.$$

Exercise 3.4.6. Prove that $\log(2), \log(3), \log(2) + \log(3), \log(2)/\log(3)$ are transcendental numbers.

Exercise 3.4.7. ★ Use the expansion

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \dots$$

to give an elementary proof that e is not a rational number (which is much easier than proving it's transcendental!).

3.4.2. *Sets defined by continued fractions.* Continued fractions allow us to define interesting sets of real numbers. For example, consider the set A of all irrational real numbers in the interval $[0, 1]$ whose 100-th partial quotient is equal to 961. This set is easily proven to be uncountable; the same is true for its complement. What is the measure of A, $\mu(A)$? Does it depend on the arbitrary value 961? Namely, would the measure be different if we asked that the 100-th partial quotient is 17? What if we asked instead that the 99-th partial quotient is 961? These are questions that we will analyze later.

Continued fractions also allows us to define interesting numbers. For example, consider the number

$$[0, 1, 2, 3, 4, \dots]$$

One can prove that this number is transcendental. In fact, as we stated before, a very general theorem says the following:

Theorem 3.4.8. *Let $m \geq 1$. Let b_0 be an integer, b_i, c_i, d_i be positive integers such that at least one of $d_1, \dots, d_m$ is not zero, then*

$$\alpha = [b_0; b_1, \dots, b_s, \overline{(c_1 + \lambda d_1, \dots, c_m + \lambda d_m)}_{\lambda=0}^{\infty}]$$

is a transcendental number. Here the notation is for blocks of natural numbers $c_1, \dots, c_m, c_1 + d_1, \dots, c_m + d_m, c_1 + 2d_1, \dots, c_m + 2d_m, c_1 + 3d_1, \dots, c_m + 3d_m, \dots$.

Combined with the next theorem, it implies that all numbers of the form

$$\frac{e^{2/y}+1}{e^{2/y}-1}, \quad y \in \mathbb{N}^+,$$

are transcendental.

Theorem 3.4.9. *We have*

$$\frac{e^{2/y}+1}{e^{2/y}-1} = [y, 3y, 5y, 7y, \dots].$$

To conform with our previous discussion of continued fractions, we take $y \in \mathbb{N}^+$ in the theorem, but in fact it holds for any positive real number y. This is a rather hard theorem; we will not prove it in this course. Note that Theorem 3.4.8 combined with Theorem 3.4.9 implies that e is transcendental: if e is algebraic, you can prove that $e^{2/y}$ is algebraic for any positive integer y and conclude that $\frac{e^{2/y}+1}{e^{2/y}-1}$ is algebraic too, and that contradicts Theorem 3.4.8. Alternately, if one is willing to assume that the continued fraction of e is

$$[2, 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, 1, 1, 10, 1, 1, 12, 1, 1, 14, 1, 1, 16, 1, 1, 18 \dots],$$

one can of course use Theorem 3.4.8 directly.

4. Approximation by rational numbers

4.1. **Types of approximations.** Let θ be a real number. The problem we consider in this chapter is to find good rational approximations to θ. Observe the following:

- If $\theta = p/q$ is rational, $\left|\theta - p/q\right| = 0$.
- If $\theta \notin \mathbb{Q}$, $\forall q \geq 1, \exists p$ such that $\left|\theta - p/q\right| < \frac{1}{q}$.

From now on we assume

$$\boxed{\theta \notin \mathbb{Q}}$$

(although, as we have illustrated before, there is some interest in getting good rational approximations for rational numbers p/q too – approximations that use a smaller denominator than q). We consider rational approximations p/q to θ that outperform these obvious ones. Below p, q denote integers such that $q \geq 1$.

BAF *Best approximations of the first kind.*
These are p, q, such that $\forall 1 \leq q' \leq q, \forall p'$, if $\frac{p}{q} \neq \frac{p'}{q'}$ then

$$\left|\theta - \frac{p'}{q'}\right| > \left|\theta - \frac{p}{q}\right|.$$

BAS *Best approximations of the second kind.*
These are p, q, such that $\forall 1 \leq q' \leq q, \forall p'$, if $\frac{p}{q} \neq \frac{p'}{q'}$ then

$$\left|q'\theta - p'\right| > \left|q\theta - p\right|.$$

OA *Optimal approximations.*
These are p, q, such that

$$\left|\theta - \frac{p}{q}\right| < \frac{1}{2q^2}.$$

Note that if p/q is a BAF then $\left|\theta - \frac{p}{q}\right| < \frac{1}{q}$, but otherwise it is hard to quantify how close it is to θ besides saying that out of all fractions with denominators at most q it is the best approximation to θ; namely, just the definition said out loud.

Lemma 4.1.1. *We have the following relations between approximations.*

(1) p/q *OA* $\implies$ p/q *BAS.*
(2) p/q *BAS* $\implies$ p/q *BAF.*

Proof. Assume that p/q is an OA and suppose that for some $q \geq d \geq 1$ and $c/d \neq p/q$ we have

$$\left|d\theta - c\right| \leq \left|q\theta - p\right|.$$

As $|q\theta - p| < 1/2q$, we have

$$\left|\theta - \frac{c}{d}\right| < \frac{1}{2dq}.$$

Then,

$$\left|\frac{c}{d} - \frac{p}{q}\right| \leq \left|\theta - \frac{c}{d}\right| + \left|\theta - \frac{p}{q}\right| < \frac{1}{2dq} + \frac{1}{2q^2} = \frac{d+q}{2dq^2}.$$

On the other hand, as $\frac{c}{d} \neq \frac{p}{q}$, $|cq - pd| \geq 1$. So,

$$\frac{1}{dq} \leq \left|\frac{c}{d} - \frac{p}{q}\right| < \frac{d+q}{2dq^2}.$$

This implies that $q < d$, which is a contradiction.

Assume now that p/q is a BAS. Given $p'/q' \neq p/q$ such that $1 \leq q' \leq q$, we have

$$|q'\theta - p'| > |q\theta - p|,$$

and so

$$\left|\theta - \frac{p'}{q'}\right| > \frac{q}{q'}\left|\theta - \frac{p}{q}\right| \geq \left|\theta - \frac{p}{q}\right|.$$

□

Exercise 4.1.2. Show that there are no inverse implications in Lemma 4.1.1.

It is interesting then to try and find such approximations. We will show that all OA and BAS arise from convergents of the continued fraction expression of θ.

Example 4.1.3. Here is an example of the approximations to $x = \frac{\sqrt{2}}{2} = 0.70710\ldots$. Using denominators up to 3 (the blue lines), the best approximation of the first kind is provided by $2/3 = 0.66666\ldots$. We need to consider fractions with denominator 7 to find a better one (red and yellow lines for denominators 4 and 5). This one is $5/7 = 0.71428\ldots$ (in grey; the diagram doesn't show fractions with denominator 6 because it is easy to see that those do not provide better approximations – the only one really in question is $5/6$ – and adding them would have made reading the diagram harder).

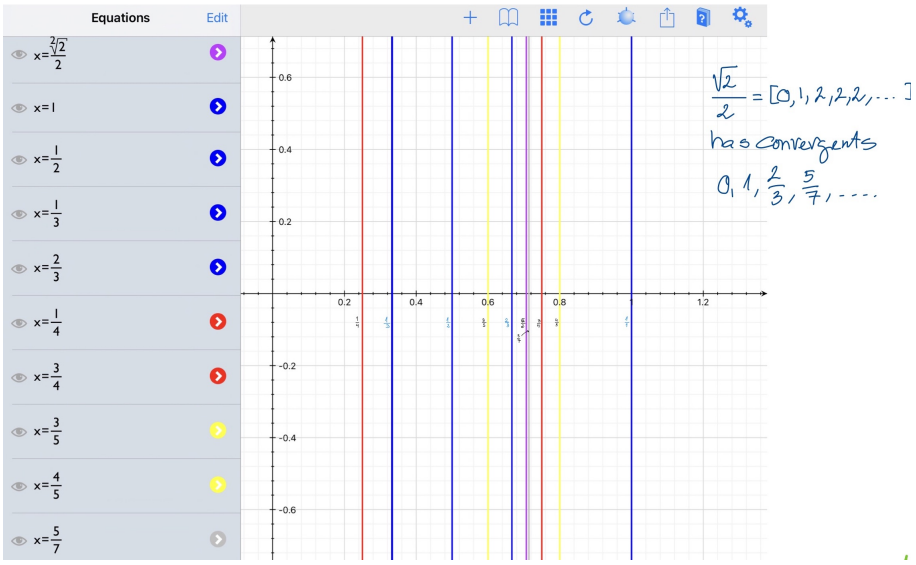
It follows from the theory we will develop that $2/3$ and $5/7$ are, in fact, even BAS, because they arise as convergents in the continued fraction expression of x. We have

$$x = \frac{\sqrt{2}}{2} = [0, 1, 2, 2, 2, 2, 2, \ldots]$$

that has convergents

$$0,\ 1,\ \frac{2}{3},\ \frac{5}{7},\ \frac{12}{17},\ \frac{29}{41},\ \frac{70}{99}\ldots.$$

The same theory would also tell us that the next BAS is 12/17 and the one after it is 29/41, and so on.



4.2. **Dirichlet's theorem.**

Theorem 4.2.1 (Dirichlet). *Let $\theta \in \mathbb{R}$. For every $Q \in \mathbb{N}, Q \geq 2$, there exists integers p, q such that $0 < q < Q$ and*

$$|q\theta - p| \leq \frac{1}{Q}.$$

Proof. Consider the following $Q+1$ numbers in the interval $[0,1]$ ("pigeons")

$$0, 1, \{\theta\}, \{2\theta\}, \ldots, \{(Q-1)\theta\}.$$

And consider the Q intervals ("pigeonholes")

$$\left[0, \frac{1}{Q}\right], \left(\frac{1}{Q}, \frac{2}{Q}\right], \ldots, \left(\frac{Q-1}{Q}, 1\right].$$

Since there are more pigeons than pigeonholes, either

- there exist $0 \leq i < j \leq Q-1$ such that $|\{i\theta\} - \{j\theta\}| \leq \frac{1}{Q}$, which implies that

$$\left|(i-j)\theta - (\lfloor i\theta \rfloor - \lfloor j\theta \rfloor)\right| \leq \frac{1}{Q},$$

  or,
- there exists $0 < i \leq Q-1$ such that $\left|\{i\theta\} - 1\right| \leq \frac{1}{Q}$ and that implies that

$$\left|i\theta - (\lfloor i\theta \rfloor + 1)\right| \leq \frac{1}{Q}.$$

□

Exercise 4.2.2. Prove that every real irrational number θ has infinitely many BAF without using Dirichlet's theorem.

Exercise 4.2.3. Prove that every real irrational number θ has infinitely many BAS (and hence also BAF) by using Dirichlet's theorem.

Corollary 4.2.4. *Assume θ is an irrational real number. There exist infinitely many q such that $(p,q) = 1$ and*

$$\left|\theta - \frac{p}{q}\right| < \frac{1}{q^2}.$$

Proof. For any $Q \in \mathbb{N}, Q \geq 2$, let p_Q, q_Q be as in Dirichlet's theorem. So, $0 < q_Q < Q$ and $|q_Q \cdot \theta - p_Q| \leq \frac{1}{Q}$. Therefore,

$$\left|\theta - \frac{p_Q}{q_Q}\right| \leq \frac{1}{q_Q Q} < \frac{1}{q_Q^2}.$$

We may assume without loss of generality that $(p_Q, q_Q) = 1$ because, if $n = \gcd(p_Q, q_Q)$, we also have

$$\left|(q_Q/n)\theta - (p_Q/n)\right| \leq \frac{1}{nQ} \leq \frac{1}{Q},$$

and so we may just replace p_Q, q_Q with $p_Q/n, q_Q/n$. Therefore, the following Claim will conclude the proof.

Claim. $\sup_Q q_Q = \infty$.

Indeed, if not, then there exists an N such that $q_Q \leq N$ for all Q. But

$$\epsilon := \min\{|q\theta - p| : q \leq N\} > 0,$$

as this minimum is essentially over a finite set (for each q there are at most 2 relevant p's for calculating the minimum; namely, $\lfloor q\theta \rfloor, \lfloor q\theta \rfloor + 1$) and equality will imply that θ is rational. Choose then Q such that $\frac{1}{Q} < \epsilon$. Then $|q_Q\theta - p_Q| \leq \frac{1}{Q} < \epsilon$. Contradiction. □

Remark 4.2.5. As you are asked to prove in Exercise 4.2.3, if θ is irrational, Dirichlet's theorem implies the existence of infinitely many BAS to θ. But even the Corollary doesn't imply that there are infinitely many optimal approximations to θ. This is indeed true, and will follow from the theory of continued fractions.

Example 4.2.6. It is certainly possible that sometimes

$$\left|\theta - \frac{p}{q}\right| < \frac{1}{q^3},$$

(or even a higher power of q). Just take $\theta = \frac{p}{q} + \frac{\sqrt{2}}{10^N}$ for sufficiently large N. But this is just a trick producing one excellent approximation for a particular θ. In general, there are powerful theorems saying that one cannot improve much on $\frac{1}{q^2}$. The most celebrated and definite result is **Roth's theorem** for which he was awarded the Fields Medal in 1958.

Theorem 4.2.7 (Roth 1955)**.** *Let θ be an irrational number. For every $\epsilon > 0$ there are only finitely many rational approximations $\frac{p}{q}, (p,q) = 1$, such that*

$$\left|\theta - \frac{p}{q}\right| < \frac{1}{q^{2+\epsilon}}.$$

Roth's theorem is a very difficult theorem; in the next section we will prove a much weaker version, Liouville's theorem that much pre-dates Roth's theorem. Liouville's theorem is very interesting nonetheless, since it is effective, unlike Roth's theorem.

4.3. **Liouville's Theorem.** .

Theorem 4.3.1 (Liouville)**.** *Let θ be a real algebraic number of degree $n > 1$.*[4] *There exists a positive constant $C = C(\theta)$ such that for all integers p, q, with $q > 0$, we have*

$$\left|\theta - \frac{p}{q}\right| \geq \frac{C}{q^n}.$$

Proof. Let $f(x)$ be the minimal polynomial of θ, which we rescale so that

$$f(x) = a_n x^n + \cdots + a_0, \quad a_i \in \mathbb{Z}, a_n \neq 0.$$

We further assume that the a_i have no proper common divisor, although this is not necessary to the proof; it just improve the constant C.

As $f(x)$ is irreducible, we have $f(p/q) \neq 0$ for all $p/q \in \mathbb{Q}$. By the mean-value theorem, for any p/q with $q > 0$ we have

$$-f(p/q) = f(\theta) - f(p/q) = (\theta - p/q) \cdot f'(\xi),$$

for a suitable ξ such that $\xi \in [\theta, p/q]$, or $[p/q, \theta]$, depending on the case.

Case 1. Suppose that $|\theta - p/q| \leq 1$. In this case, find a constant C_1 such that

$$|f'(\xi)| \leq C_1, \quad \forall \xi \in [\theta - 1, \theta + 1].$$

Then, $|\theta - p/q| \geq \frac{1}{f'(\xi)} \cdot |f(p/q)| \geq \frac{1}{C_1} \cdot |f(p/q)|$. But,

$$f(p/q) = (a_n p^n + \cdots + a_1 p q^{n-1} + a_0 q^n)/q^n = (\text{non-zero integer})/q^n.$$

Therefore,

$$|\theta - p/q| \geq \frac{1}{C_1} \cdot \frac{1}{q^n}.$$

Case 2. $|\theta - p/q| \geq 1$. Then $|\theta - p/q| \geq \frac{1}{q^n}$.

Let then

$$C = \min\{\frac{1}{C_1}, 1\}.$$

□

Exercise 4.3.2. Analyze the proof of Liouville's theorem and find a constant C as in the theorem for $\sqrt{2}, \frac{1+\sqrt{5}}{2}, \sqrt[3]{5} \in \mathbb{R}$.

Exercise 4.3.3. Let $\theta = \sum_{n=1}^{\infty} \frac{1}{10^{n!}}$. Prove that θ is transcendental. This application of Liouville's theorem was given by him in 1844 and produced the first explicitly given number that was provenly transcendental.

Exercise 4.3.4. ★ Construct a set T of real transcendental numbers with $\left|T\right| > \aleph_0$, and $\mu(T) = 0$.

[4] That is, θ is algebraic, but not rational.

5. Continued fractions and approximations

We will assume throughout this section that θ is an irrational real number. Some of the statements can be extended, or generalized to rational numbers, but we will not do so here.

Recall that we have already proved that for θ a real irrational number we have

$$\left|\theta - \frac{p_n}{q_n}\right| < \frac{1}{q_n^2},$$

where $\{p_n/q_n\}$ are the convergents to θ. Let us improve on that.

Theorem 5.0.1. *For any $n \geq 0$, either $\frac{p_n}{q_n}$ or $\frac{p_{n+1}}{q_{n+1}}$ is an optimal approximation.*

Proof. If not, then for some n we have both

$$\left|\theta - \frac{p_n}{q_n}\right| \geq \frac{1}{2q_n^2}, \quad \left|\theta - \frac{p_{n+1}}{q_{n+1}}\right| \geq \frac{1}{2q_{n+1}^2}.$$

As θ lies *between* p_n/q_n and p_{n+1}/q_{n+1} (Corollary 2.1.9), we conclude that

$$\left|\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n}\right| \geq \frac{1}{2}\left(\frac{1}{q_n^2} + \frac{1}{q_{n+1}^2}\right).$$

On the other hand,

$$\left|\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n}\right| = \frac{1}{q_n q_{n+1}}.$$

It follows that

$$\frac{1}{q_n q_{n+1}} \geq \frac{1}{2}\left(\frac{1}{q_n^2} + \frac{1}{q_{n+1}^2}\right).$$

Multiply both sides by $2q_n^2 q_{n+1}^2$ and rearrange to conclude that $0 \geq (q_{n+1} - q_n)^2$. This implies that $q_{n+1} = q_n$. However, we proved in Corollary 2.1.6 that q_n is strictly monotone increasing for $n \geq 1$. So the only case we still need to consider is when $n = 0$, and $q_0 = 1, q_1 = 1 = a_1$. In this case, if $\frac{p_0}{q_0} = \lfloor\theta\rfloor = a_0$ is not an optimal approximation then $\{\theta\} \geq \frac{1}{2}$, in fact $\{\theta\} > \frac{1}{2}$ because θ is irrational. But then

$$0 < \frac{p_1}{q_1} - \theta = \frac{a_0 + 1}{1} - (\lfloor\theta\rfloor + \{\theta\}) = 1 - \{\theta\} < \frac{1}{2}.$$

That means that $\frac{p_1}{q_1}$ is an optimal approximation. □

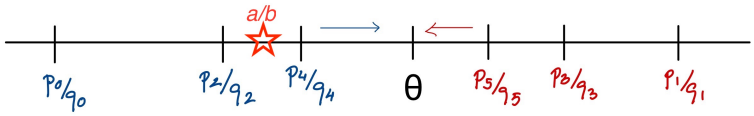
As every OA is a BAS, we conclude the following.

Corollary 5.0.2. *For all n, either $\frac{p_n}{q_n}$ or $\frac{p_{n+1}}{q_{n+1}}$ is a BAS for θ.*

In fact, we will soon prove that *all* $\frac{p_n}{q_n}$ are BAS (with essentially one exception). But first, let's prove that all *BAS* arise as convergents.

Theorem 5.0.3. *Let θ is an irrational real number. Any BAS for θ is a convergent $\frac{p_n}{q_n}$ to θ.*

Proof. Recall the picture of convergence from Corollary 2.1.9:



Let a/b be a BAS for θ. We will first show that it lies between p_0/q_0 and p_1/q_1.

If $\frac{a}{b} < a_0 = \frac{p_0}{q_0}$ then $|1 \cdot \theta - a_0| < |\theta - \frac{a}{b}| \leq |b\theta - a|$. So, $a_0/1$ "beats" a/b. Namely, a/b cannot be a BAS. Similarly, if $\frac{a}{b} > \frac{p_1}{q_1}$ then $|\theta - \frac{a}{b}| > |\frac{p_1}{q_1} - \frac{a}{b}| = |\frac{bp_1 - aq_1}{bq_1}| \geq \frac{1}{bq_1}$. Multiplying by b we conclude that

$$|b\theta - a| > \frac{1}{q_1} = \frac{1}{a_1} \geq |\theta - a_0|.$$

The last inequality follows from the fact that $\theta - a_0 = \frac{1}{a_1+}\rfloor\frac{1}{a_2+}\rfloor \ldots \frac{1}{a_{N-1}+}\rfloor \ldots$ Thus, again we find that $a_0/1$ beats a/b and so a/b cannot be a BAS. Contradiction.

Let assume that a/b is not a convergent to θ. Then a/b is strictly between $\frac{p_{k-1}}{q_{k-1}}$ and $\frac{p_{k+1}}{q_{k+1}}$ for some $k \geq 0$. We will derive a contradiction by showing that p_k/q_k beats a/b. For that we should first establish that $q_k \leq b$. Note that, on the one hand, we have

$$\left|\frac{a}{b} - \frac{p_{k-1}}{q_{k-1}}\right| = \left|\frac{aq_{k_1} - bp_{k-1}}{bq_{k-1}}\right| \geq \frac{1}{bq_{k-1}}.$$

On the other hand, by Corollary 2.1.5, we have

$$\left|\frac{a}{b} - \frac{p_{k-1}}{q_{k-1}}\right| < \left|\frac{p_k}{q_k} - \frac{p_{k-1}}{q_{k-1}}\right| = \frac{1}{q_k q_{k-1}}.$$

It follows that $q_k < b$.

Now, because a/b is between $\frac{p_{k-1}}{q_{k-1}}$ and $\frac{p_{k+1}}{q_{k+1}}$,

$$\left|\theta - \frac{a}{b}\right| \geq \left|\frac{p_{k+1}}{q_{k+1}} - \frac{a}{b}\right| \geq \frac{1}{bq_{k+1}}.$$

Therefore,

$$\left|b\theta - a\right| \geq \frac{1}{q_{k+1}}.$$

But, by Corollary 2.1.10,

$$\left|q_k\theta - p_k\right| < \frac{1}{q_{k+1}}.$$

And we got a contradiction: "p_k/q_k beats a/b". □

Theorem 5.0.4. *Let θ be an irrational real number. Any convergent $\frac{p_k}{q_k}$ to θ is a BAS for θ with the only possible exception being $\frac{p_0}{q_0}$.* [5]

Proof. Let $k \geq 1$. For $x \in \mathbb{Z}, y \in \{1, 2, \ldots, q_k\}$, consider

$$\min_{x,y} |y\theta - x|.$$

Choose a minimal y_0 for which this minimum is achieved. For that y_0 there is a unique x_0 such that $|y_0\theta - x_0|$ is the minimum, else θ is rational. Therefore, by definition, $\frac{x_0}{y_0}$ is a BAS and by Theorem 5.0.3 there exists an s such that

$$\frac{x_0}{y_0} = \frac{p_s}{q_s}.$$

Now, as $y_0 \leq q_k$ and the q_k are monotone increasing, we must have $s \leq k$. If $s = k$, we are done (make sure you understand why!). So, assume $s < k$. Using Theorem 2.3.3, we have

$$\left|q_s\theta - p_s\right| > \frac{1}{q_{s+1} + q_s} \geq \frac{1}{q_k + q_{k-1}}.$$

[5] There is a mistake in Khinchin in that the exception is not stated. Clearly, whenever $\{\theta\} > 1/2$, $p_0/q_0 = \lfloor\theta\rfloor$ is not a BAS; in this case $q_1 = 1$ and $p_1/q_1 = \lfloor\theta\rfloor + 1$ is better.

On the other hand, by Corollary 2.1.10,

$$\left|q_k\theta - p_k\right| < \frac{1}{q_{k+1}}.$$

As, by the definition of p_s, q_s, $|q_s\theta - p_s| \leq |p_k\theta - q_k|$, we can combine the two inequalities and find that

$$\frac{1}{q_{k+1}} > \frac{1}{q_k + q_{k-1}}.$$

This implies $q_{k+1} < q_k + q_{k-1}$ and that is a contradiction as $q_{k+1} = a_{k+1}q_k + q_{k-1}$ and all quantities appearing in this recursion formula are positive integers.

□

Let us summarize all we have learned about the relation between approximations for an irrational real number θ and its convergents $\frac{p_n}{q_n}$.

- For every n, either $\frac{p_n}{q_n}$ or $\frac{p_{n+1}}{q_{n+1}}$ is an OA.
- Every convergent $\frac{p_n}{q_n}, n \geq 1$ is a BAS.
- Every BAS for θ is a convergent to θ.
- In general, θ will have many BAF that aren't BAS and so do not arise from the convergents to θ.

Remark 5.0.5. One can show that all BAFs for θ can be derived from the continued fraction expansion as well by "suitably combining" the partial quotients. We don't discuss that here, but this result can be found, for example, in Khinchin's book.

6. Quadratic irrationals, Pell's equation and continued fractions

In this section we investigate periodic continued fractions. Where by "periodic" we really mean ultimately periodic. That is, a **periodic continued fraction** is a continued fraction of the form

$$[a_0, \dots, a_{k_0-1}, b_1, \dots, b_h, b_1, \dots, b_h, b_1, \dots, b_h, \dots],$$

where $a_0 \in \mathbb{Z}, a_i, b_i \in \mathbb{N}^+$, for $i > 0$. A common notation is

$$[a_0, \dots, a_{k_0-1}, \overline{b_1, \dots, b_h}]$$

We will study such continued fractions and connect them to solutions to **Pell's equation**. Pell's equation is an equation of the form

$$x^2 - dy^2 = 1, \tag{2}$$

where d is a positive integer that is not a square. We will not prove every theorem we present in this section. The proofs can be found in many other textbooks in number theory. In general, we tend to omit proofs that are either way too difficult (like for Roth's, or Baker's, theorem) or that are very technical and long, involving analysis of multiple cases.

6.1. **Periodic continued fractions.**

Proposition 6.1.1. *Let θ be a periodic continued fraction,*

$$[a_0, \ldots, a_{k_0-1}, \overline{b_1, \ldots, b_h}].$$

Then θ is quadratic over $\mathbb{Q}$. Namely, it is an algebraic number of degree 2.

Proof. As θ is irrational, we only need to show that θ satisfies a quadratic equation with rational coefficients. Let r be the (irrational) number,

$$r = [\overline{b_1, \ldots, b_h}].$$

Then,

$$\theta = [a_0, \ldots, a_{k_0-1}, r] = [a_0, \ldots, a_{k_0-1}, b_1, \ldots, b_h, r].$$

Letting p_a, q_a denote the convergents to θ, we find that

$$\theta = \frac{rp_{k_0-1} + p_{k_0-2}}{rq_{k_0-1} + q_{k_0-2}} = \frac{rp_{k_0+h-1} + p_{k_0+h-2}}{rq_{k_0+h-1} + q_{k_0+h-2}}.$$

Clearing denominators and rearranging, it follows that r satisfies a quadratic equation over $\mathbb{Q}$ and so θ satisfies one too. □

Theorem 6.1.2. *If θ satisfies a quadratic equation $ax^2 + bx + c, a \neq 0$, over $\mathbb{Q}$ then θ has an ultimately periodic continued fraction expansion $[a_0, \ldots, a_{k_0-1}, \overline{b_1, \ldots, b_h}]$.*

The proof is not very difficult, but we will omit it for the reasons mentioned above. It can be found in Khinchin's book pp. 48 - 50.

Here are some examples of quadratic irrational numbers and their expansions.

$$(1 + \sqrt{5})/2 = [\overline{1}], \quad 1 + \sqrt{2} = [\overline{2}], \quad 3 + 2\sqrt{2} = [5, \overline{1, 4}].$$

In general, things can behave rather unexpectedly. For example,

$$\frac{5 + \sqrt{2}}{11} = [0, 1, 1, 2, \ \overline{1, 1, 30, 1, 1, 3, 1, 14, 1, 3}\ ].$$

The length of the period of $\sqrt{d}$ varies erratically,

$\sqrt{39}$ $[6, \overline{4, 12}]$

$\sqrt{40}$ $[6, \overline{3, 12}]$

$\sqrt{41}$ $[6, \overline{2, 2, 12}]$

$\sqrt{42}$ $[6, \overline{2, 12}]$

$\sqrt{43}$ $[6, \overline{1, 1, 3, 1, 5, 1, 3, 1, 1, 12}]$

$\sqrt{44}$ $[6, \overline{1, 1, 1, 2, 1, 1, 1, 12}]$

$\sqrt{45}$ $[6, \overline{1, 2, 2, 2, 1, 12}]$

$\sqrt{46}$ $[6, \overline{1, 3, 1, 1, 2, 6, 2, 1, 1, 3, 1, 12}]$

$\sqrt{47}$ $[6, \overline{1, 5, 1, 12}]$

$\sqrt{48}$ $[6, \overline{1, 12}]$

Using the command in PARI

```
for(n=8,60, print(n"^(1/2)", "=", contfrac(sqrt(n), 30)))
```

you will get as output the continued fraction expansion of $\sqrt{n}$ for all integers n between 8 and 60.

Remark 6.1.3. The length of a period is a rather intriguing quantity on which there is a lot to say; for example, under the generalized Riemann hypothesis, one can prove that the length is $O(\sqrt{n}\log\log n)$. That is, there is a constant C such that for $x \gg 0$, the length is at most $C\sqrt{n}\log\log n$. This result is not so easy to prove. In contrast, it is easy to prove that $\sqrt{n}$ has an expression as a continued fraction of the form $[a,\overline{b}\,]$ if and only if $b = 2a$, in which case $n = a^2+1$. If we use also Theorem 6.2.4 below that gives us some information of the general form of the development of $\sqrt{n}$, for n not a square, we find that the only square roots $\sqrt{n}$ with period of length 1 are those of the form $\sqrt{a^2+1}$ and their continued fraction expansion is

$$\sqrt{a^2+1} = [a,\overline{2a}\,].$$

In particular, we deduce that there are arbitrarily large n such that $\sqrt{n}$ has a period of length 1.

6.2. **Pell's equation.** Let d be a positive integer that is not a square. Consider Pell's equation.

$$x^2 - dy^2 = 1. \tag{3}$$

One is interested in solutions in integers for such an equation. Pell equations may be considered as among the simplest of diophantine equations, and this is one reason to seek a thorough understanding of them; at the same time, the solutions are numbers of the form $a+b\sqrt{d}$, where a,b are integers. As a result $a+b\sqrt{d} \in \mathbb{Q}(\sqrt{d})$ is a unit of the ring of algebraic integers of $\mathbb{Q}(\sqrt{d})$.[6]

The behaviour of the **fundamental solution**, namely, of the smallest pair of positive integers (x,y) solving the equation - smallest in the sense that any other such pair (x',y') satisfies $y' > y$ - is mysterious for the same reason the length of the period of the continued fraction of $\sqrt{d}$ is mysterious. To illustrate the point, the fundamental solution to the equation

$$x^2 - 1140y^2 = 1$$

is

$$x = 2431, \quad y = 72,$$

and the period of $\sqrt{1140}$ is 6. In contrast, the fundamental solution of

$$x^2 - 1141y^2 = 1$$

is

$$x = 1036782394157223963237125215, \quad , y = 30693385322765657197397208,$$

and the length of the period of $\sqrt{1141}$ is 58.

The following theorem connects between integer solutions for Pell's equations and continued fractions. In fact, it applies to somewhat more general equations.

Theorem 6.2.1. *Let $0 < N < \sqrt{d}$ be an integer, where d is not a square. Let s,t be positive integer solutions to the equation*

$$x^2 - dy^2 = N,$$

with $\gcd(s,t) = 1$. Then, for some n, we have

$$\frac{s}{t} = \frac{p_n}{q_n},$$

where p_n/q_n is a convergent to $\sqrt{d}$.

[6] An algebraic number α is called an algebraic integer if α solves a monic polynomial with integer coefficients. The algebraic numbers form a subring of $\bar{\mathbb{Q}}$.

Proof. The idea is to show that s/t is an OA for $\sqrt{d}$, hence a BAS. The theorem follows then from Theorem 5.0.3.

As $s^2 - dt^2 = N$ we have $(\frac{s}{t} + \sqrt{d})(\frac{s}{t} - \sqrt{d}) = \frac{N}{t^2}$ and so

$$\frac{s}{t} - \sqrt{d} = \frac{N}{t^2(\frac{s}{t} + \sqrt{d})} < \frac{1}{t^2(\frac{s}{t\sqrt{d}} + 1)}$$

(as $N < \sqrt{d}$). On the other hand, as $\frac{s}{t} - \sqrt{d} > 0$, it follows that $\frac{s}{t\sqrt{d}} > 1$ and so

$$\left|\sqrt{d} - \frac{s}{t}\right| < \frac{1}{2t^2}.$$

This proves that s/t is an OA for $\sqrt{d}$. □

We record a particular case:

Corollary 6.2.2. *Any positive solution to Pell's equation*

$$x^2 - dy^2 = 1$$

arises as a convergent to $\sqrt{d}$.

Note that the Corollary does not claim that every convergent is a solution to Pell's equation, and, in fact, it's not true. Let us look at an example.

Example 6.2.3. $\sqrt{7} = [2, \overline{1, 1, 1, 4}\,]$. We have the following convergents

$$\frac{p_0}{q_0} = \frac{2}{1}, \quad \frac{p_1}{q_1} = \frac{3}{1}, \quad \frac{p_2}{q_2} = \frac{5}{2}, \quad \frac{p_3}{q_3} = \frac{8}{3}, \quad \frac{p_4}{q_4} = \frac{37}{14}, \quad \frac{p_5}{q_5} = \frac{45}{17}, \ldots$$

Then, correspondingly, p_i/q_i is a solution for the following equations:

$$x^2 - 7y^2 = -3, \quad x^2 - 7y^2 = 2, \quad x^2 - 7y^2 = -3, \quad x^2 - 7y^2 = 1, \quad x^2 - 7y^2 = -3, \quad x^2 - 7y^2 = 2.$$

As is clear from this example, the complete story as to what Pell equation-like the convergents solve is rather intricate. We don't give here the full story but only one theorem, without proof.

Theorem 6.2.4. *Suppose that* $d > 0$ *and is not a square. Then*

$$\sqrt{d} = [a_0, \overline{b_1, \ldots, b_n}\,].$$

- *If* n *is* <u>*even*</u>*, the positive solutions to* $x^2 - dy^2 = 1$ *are*

$$(p_{jn-1}, q_{jn-1}), \quad j = 1, 2, 3, \ldots$$

- *If* n *is* <u>*odd*</u>*, the positive solutions to* $x^2 - dy^2 = 1$ *are*

$$(p_{2jn-1}, q_{2jn-1}), \quad j = 1, 2, 3, \ldots$$

Exercise 6.2.5. Find positive solutions for the following equations:

(1) $x^2 - 39y^2 = 1$.
(2) $x^2 - 41y^2 = 1$.

Exercise 6.2.6. Prove that there are infinitely many positive solutions to the equation

$$x^2 - 39y^2 = -3.$$

(Hint: given a solution (a, b) to $x^2 - 39y^2 = -3$ and a solution (c, d) to $x^2 - 39y^2 = 1$, show that one can generate a new solution to $x^2 - 39y^2 = -3$ by using the product $(a + b\sqrt{39})(c + d\sqrt{39})$.)

Exercise 6.2.7. Find a positive solution to the equation $x^2 - 41y^2 = 5$.

Exercise 6.2.8. ★ **Triangular numbers** are the integers $1, 3, 6, \ldots, \frac{n(n+1)}{2}, \ldots$.

Show that there are infinitely many triangular numbers that are squares and find 3 of them besides $0, 1$.[7]

Exercise 6.2.9. ★ Find five pairs of integers (n, N), $1 \leq n \leq N$, such that

$$1 + 2 + \cdots + (n-1) = (n+1) + (n+2) + \cdots + N.$$

Exercise 6.2.10. Let (a, b) be a solution to Pell's equation $x^2 - dy^2 = 1$. Show that for any n, if we define A_n, B_n as follows

$$A_n + B_n\sqrt{d} = (a + b\sqrt{d})^n,$$

then A_n, B_n are also solutions to the same equation. Use this to show that if a Pell equation $x^2 - dy^2 = N$ has a solution then it has infinitely many solutions.

Exercise 6.2.11. ★ Show that there are infinitely many solutions (a, b) to $x^2 - 10y^2 = 1$ such that $7|a$.

Exercise 6.2.12. The equation $x^2 - dy^2 = -1$ doesn't always have integral solutions: prove that if $d \equiv 0, -1 \pmod 4$ there are no integral solutions. However, prove that if a solution exists then it is a convergent to $\sqrt{d}$.

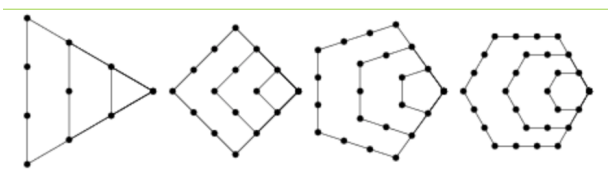
Exercise 6.2.13. Prove that $[a, \overline{b, c}] = \sqrt{n}$ for some positive integer n, if and only if $a > 0$, $c = 2a$ and $b|c$, in which case $n = a^2 + c/b$. Deduce that there are arbitrarily large n such that $\sqrt{n}$ has a period of length 2. Compare with Remark 6.1.3.

Remark 6.2.14. One can prove that if a solution to $x^2 - dy^2 = -1$ exists, then the length of the period of $\sqrt{d}$ must be odd. And in fact a theorem similar to Theorem 6.2.4 holds. There is no known criterion to determine when this happens. This is known as the question of "the sign of the fundamental unit", because it has to do with the question whether there is a unit in the ring of integers of $\mathbb{Q}(\sqrt{d})$ that has norm -1 to $\mathbb{Q}$.

7. The measure of some sets defined by continued fractions

We change gears in this section. Our purpose is to look at sets of real numbers defined by properties of continued fractions and ask how "big" they are. More precisely, we will look at sets contained in $[0, 1]$ – just for convenience, the generalization is easy – that are defined in terms

[7] There is the more general notion of figurative numbers, or k-gonal numbers. The 3-gonal numbers are the triangular numbers $\{n(n+1)/2 : n \in \mathbb{N}^+\}$. The 4-gonal numbers are the squares $\{n^2 : n \in \mathbb{N}^+\}$. Similarly, the pentagonal numbers are given by $\{n(3n-1)/2 : n \in \mathbb{N}^+\}$. The k-gonal numbers are $\{\frac{k}{2}(n^2 - n) - n^2 + 2n : n \in \mathbb{N}^+\}$.



It was discovered by Fermat in 1636, and proved first by Cauchy in 1813, that every positive integer is a sum of k k-gonal numbers. For example, a sum of 3 triangular numbers, a sum of 4 squares and so on.

of properties of their continued fractions and ask for their Lebesgue measure. As an example, we may consider the set

$$S = \{[0, a_1, a_2, \dots] : a_1 = 2, a_2 = 3\},$$

and ask for $\mu(S)$. This is quite easy to calculate (try!), but already finding the measure of the set

$$T = \{[0, a_1, a_2, \dots] : a_1 = a_2\},$$

requires more thought (try this one too!). So one wants to develop some general methodology.

One of the main ideas is to think of $a_1, a_2, \dots$ as functions of x. For every $x \in [0, 1]$ we can write

$$x = [0, a_1, a_2, \dots],$$

for uniquely determined positive integers a_i that depend on x (unless x is rational, but those have measure 0, so we disregard them). So,

$$x = [0, a_1(x), a_2(x), \dots],$$

and for every $k \geq 1$ we have the functions

$$a_k \colon [0, 1] \to \mathbb{N}^+, \quad x \mapsto a_k(x),$$

which we want to understand.

7.1. **The functions $a_i(x)$.** Before proving the lemma we need, let us look at the functions $a_1(x)$ and $a_2(x)$ to get some feeling as to the general behaviour. We wish to understand, for $x \in [0, 1]$ when does $a_1(x) = k_1$, where k_1 is some positive integer. As ultimately we want to analyze all functions $a_i(x)$, this really makes sense only for $x \notin \mathbb{Q}$. And so we will usually assume that.

So, the question is what is the location of all $x \in [0, 1]$, or $x \in [0, 1] \setminus \mathbb{Q}$ to be pedantic, such that $a_1(x) = k_1$. Such an x is then written as

$$x = [0, k_1, r],$$

where $1 < r < \infty$ (the strict inequality because x is irrational). Thus,

$$x = \frac{1}{k_1 + \frac{1}{r}}.$$

As r ranges over $(1, \infty)$, $\frac{1}{r}$ ranges over $(0, 1)$. So $x \in (\frac{1}{k_1+1}, \frac{1}{k_1})$. Note that $\frac{p_0}{q_0} = \frac{0}{1}$, $\frac{p_1}{q_1} = \frac{1}{k_1}$ and so we can write

$$\left(\frac{1}{k_1 + 1}, \frac{1}{k_1}\right) = \left(\frac{p_1 + p_0}{q_1 + q_0}, \frac{p_1}{q_1}\right) = \{x : x = [0, k_1, a_2, a_3, \dots] : a_i \in \mathbb{N}^+\},$$

where $\frac{p_0}{q_0}, \frac{p_1}{q_1}$ are those of any x in $\{x : x = [0, k_1, a_2, a_3 \dots]\}$.

The function $a_1(x)$ therefore has the graph appearing in Figure 1 (where, again, rationality allows us not to bother with interval ends).

What about $a_2(x)$ then? Suppose that $a_2(x) = k_2$. It is easier to analyze the situation after giving some definite value to $a_1(x)$. So, suppose that $a_1(x) = k_1$. Then, we want to know the location of all x of the form

$$x = [0, k_1, k_2, r], \quad r \in (1, \infty).$$

Recall that using modified p_n/q_n we can write x as

$$x = \frac{rp_2 + p_1}{rq_2 + q_1} = \frac{p_2 + p_1/r}{q_2 + q_1/r}.$$

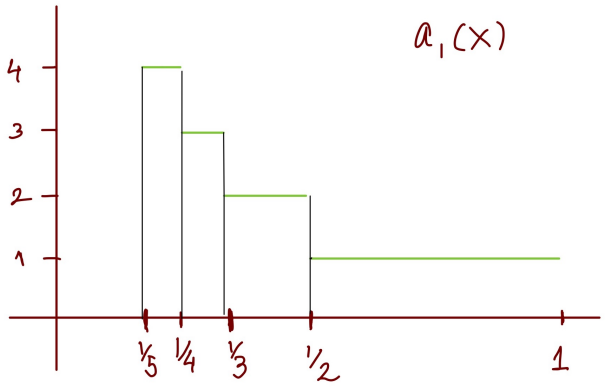


FIGURE 1. The function $a_1(x)$

When $r = 1$ we get $\frac{p_2+p_1}{q_2+q_1}$ and when $r = \infty$ we get $\frac{p_2}{q_2}$. Thus, and we will be more rigorous when we prove the general statement, we have

$$\left(\frac{p_2}{q_2}, \frac{p_2+p_1}{q_2+q_1}\right) = \{x : x = [0, k_1, k_2, a_3, a_4, ...], a_i \in \mathbb{N}^+\},$$

where $\frac{p_1}{q_1}, \frac{p_2}{q_2}$ are those of any x in $\{x : x = [0, k_1, k_2, a_3, a_4, \dots]\}$.

Moreover, $p_1 = 1, p_2 = k_2, q_1 = k_1, q_2 = k_1k_2 + 1$, so

$$\left(\frac{k_2}{k_1k_2+1}, \frac{k_2+1}{k_1k_2+1+k_1}\right) = \{x : x = [0, k_1, k_2, a_3, a_4, ...]\}.$$

Note that when $k_2 = 1$ we get the end point $\frac{k_2}{k_1k_2+1} = \frac{1}{k_1+1}$ and as k_2 goes to infinity, $\frac{k_2+1}{k_1k_2+1+k_1} \to \frac{1}{k_1}$. That is, every interval

$$\left(\frac{1}{k_1+1}, \frac{1}{k_1}\right)$$

on which $a_1(x) = k_1$, is divided into a disjoint union of intervals

$$\left(\frac{k_2}{k_1k_2+1}, \frac{k_2+1}{k_1(k_2+1)+1}\right),$$

where on such an interval $a_2(x) = k_2$. Otherwise said:

Over every step of $a_1(x)$, $a_2(x)$ is a step function. See Figure 2.

Exercise 7.1.1. Find $\mu(S)$ and $\mu(T)$ where

$$S = \{[0, a_1, a_2, \dots] : a_1 = 2, a_2 = 3\}, \quad T = \{[0, a_1, a_2, \dots] : a_1 = a_2\}.$$

For T, the answer should be expressed as an infinite sum to which you should provide non-trivial lower and upper bounds (say, different than 0 or 1).

Let us introduce the following notation. For integers $1 \leq n_1 < n_2 < \cdots < n_s$ and any positive integers $k_1, \dots, k_s$, let

$$E\left(\begin{smallmatrix} n_1 & n_2 & \dots & n_s \\ k_1 & k_2 & \dots & k_s \end{smallmatrix}\right) = \{x \in (0,1) : a_{n_i}(x) = k_i, i = 1, 2, \dots, s\}.$$

For example,

$$E\left(\begin{smallmatrix} 1 & 2 & \dots & s \\ k_1 & k_2 & \dots & k_s \end{smallmatrix}\right) = \{x = [0, k_1, k_2, \dots, k_s, a_{s+1}, \dots]\}.$$

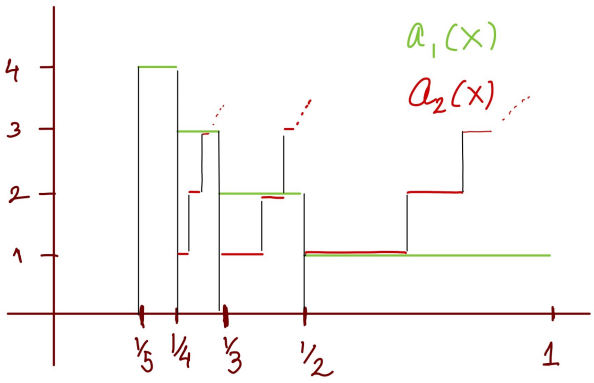


FIGURE 2. The function $a_2(x)$

In these terms, we see that what we have analyzed above were $E\left(\begin{smallmatrix}1\\k_1\end{smallmatrix}\right)$ and $E\left(\begin{smallmatrix}1&2\\k_1&k_2\end{smallmatrix}\right)$.

To understand the following Lemma, it is useful to note that given two rational numbers $\frac{a}{b} < \frac{c}{d}$ with positive denominators, we have

$$\frac{a}{b} < \frac{a+c}{b+d} < \frac{c}{d}.$$

Lemma 7.1.2. *$E\left(\begin{smallmatrix}1&2&\cdots&n\\k_1&k_2&\cdots&k_n\end{smallmatrix}\right)$ is an interval with end points $\frac{p_n+p_{n-1}}{q_n+q_{n-1}}$ and $\frac{p_n}{q_n}$ (which is bigger than which depends on the parity of n), where p_i/q_i, for $i \leq n$, are the partial quotients of any $x \in E\left(\begin{smallmatrix}1&2&\cdots&n\\k_1&k_2&\cdots&k_n\end{smallmatrix}\right)$.*

Proof. The proof is by induction, where we have already checked the cases $n = 1, 2$. So, assume that $E\left(\begin{smallmatrix}1&2&\cdots&n\\k_1&k_2&\cdots&k_n\end{smallmatrix}\right)$ is an interval with end points $\frac{p_n+p_{n-1}}{q_n+q_{n-1}}$ and $\frac{p_n}{q_n}$.

We have

$$E\left(\begin{smallmatrix}1&2&\cdots&n&n+1\\k_1&k_2&\cdots&k_n&k_{n+1}\end{smallmatrix}\right) = \{x = [0, k_1, k_2, \ldots, k_n, k_{n+1}, r_{n+2}] : 1 < r_{n+2} < \infty\}$$

(where, as usual, we ignore the case $r_{n+2} = 1$ as it corresponds to a rational number). Thus, much as we have done for a_2, we find that

$$x = \frac{r_{n+2}p_{n+1} + p_n}{r_{n+2}q_{n+1} + q_n} = \frac{p_{n+1} + p_n/r_{n+2}}{q_{n+1} + q_n/r_{n+2}}.$$

In this expression, p_{n+1}, p_n, q_{n+1} and q_n are fixed by the data $k_1, k_2, \ldots, k_{n+1}$ and we view the expression for x as a function of r. The limit as r goes to 1 is $\frac{p_{n+1}+p_n}{q_{n+1}+q_n}$ and when r goes to infinity, $\frac{p_{n+1}}{q_{n+1}}$. Moreover, the derivative as a function of r is

$$\frac{p_{n+1}(r_{n+2}q_{n+1} + q_n) - q_{n+1}(r_{n+2}p_{n+1} + p_n)}{(r_{n+2}q_{n+1} + q_n)^2} = \frac{(-1)^n}{(r_{n+2}q_{n+1} + q_n)^2}.$$

So $\frac{r_{n+2}p_{n+1}+p_n}{r_{n+2}q_{n+1}+q_n}$ is a monotone function we conclude that as r varies from 1 to ∞, the x that we get cover an interval with end points $\frac{p_{n+1}+p_n}{q_{n+1}+q_n}$ and $\frac{p_{n+1}}{q_{n+1}}$. (If n is even this is the interval $(\frac{p_{n+1}+p_n}{q_{n+1}+q_n}, \frac{p_{n+1}}{q_{n+1}})$ and if n is odd this is the interval $(\frac{p_{n+1}}{q_{n+1}}, \frac{p_{n+1}+p_n}{q_{n+1}+q_n})$.) □

Lemma 7.1.3. *We have*

$$\rho := \frac{\mu\left(E\left(\begin{smallmatrix}1&2&\cdots&n&n+1\\k_1&k_2&\cdots&k_n&s\end{smallmatrix}\right)\right)}{\mu\left(E\left(\begin{smallmatrix}1&2&\cdots&n\\k_1&k_2&\cdots&k_n\end{smallmatrix}\right)\right)} = \frac{1}{s^2} \cdot \frac{1 + \frac{q_{n-1}}{q_n}}{(1 + \frac{q_{n-1}}{sq_n})(1 + \frac{1}{s} + \frac{q_{n-1}}{sq_n})};$$

It satisfies,

$$\frac{1}{3s^2} < \rho < \frac{2}{s^2},$$

independently of $k_1, \dots, k_n$ *(!)*

Exercise 7.1.4. Prove Lemma 7.1.3. (The expressions look daunting at first, but remember that we know to describe these sets using intervals. That, after algebraic manipulations, provides the first claim. For one of the inequalities, simply replace $1 + \frac{q_{n-1}}{q_n}$ by $1 + \frac{q_{n-1}}{sq_n}$). Can you improve the constants 2 and $\frac{1}{3}$?

Corollary 7.1.5. *We have*

$$\frac{1}{3s^2} < \mu(E\left(\begin{smallmatrix} n+1 \\ s \end{smallmatrix}\right)) < \frac{2}{s^2}.$$

Proof. We have

$$\mu((0,1)) = \sum_{(k_1,\dots,k_n)} \mu\left(E\left(\begin{smallmatrix} 1 & 2 & \dots & n \\ k_1 & k_2 & \dots & k_n \end{smallmatrix}\right)\right),$$

while

$$\mu(E\left(\begin{smallmatrix} n+1 \\ s \end{smallmatrix}\right)) = \sum_{(k_1,\dots,k_n)} \mu\left(E\left(\begin{smallmatrix} 1 & 2 & \dots & n & n+1 \\ k_1 & k_2 & \dots & k_n & s \end{smallmatrix}\right)\right).$$

Comparing like terms, and using Lemma 7.1.3, the Corollary follows. □

So, while it is not true that $\mu(E\left(\begin{smallmatrix} n+1 \\ s \end{smallmatrix}\right))$ is independent of n, it is closed to being so. For example, try the following.

Exercise 7.1.6. Prove that $\mu(E\left(\begin{smallmatrix} 1 \\ 2 \end{smallmatrix}\right)) = \frac{1}{6} = 0.166...$, while

$$\mu(E\left(\begin{smallmatrix} 2 \\ s \end{smallmatrix}\right)) = \sum_{k=1}^{\infty} \frac{1}{(2k+1)(3k+1)}.$$

The value of this series is numerically close to 0.1685. How well can you approximate this sum?

7.2. **Numbers with bounded partial quotients.** In this section we consider a set S of real numbers S contained in $[0,1]$ that exhibits a number of different characteristics. On the one hand, it's a "small" set, since it the set of all real numbers in $(0,1)$ whose continued fraction expansion involves only finitely many integers. It is also the set of real numbers that cannot be approximated too well by rational numbers, in a sense defined below (see Theorem 7.2.4). From a different angle, it is a "huge set" since it has the same cardinality, $2^{\aleph_0}$, as $\mathbb{R}$. And, still from a different perspective, we will prove later that it has Hausdorff dimension 1. Each of these facts brings to light another aspect of this set.

Theorem 7.2.1. *Let*

$$S = \{x = [0, a_1, a_2, \dots] : \exists M = M(x) \text{ such that } a_i \leq M, \forall i\}.$$

This is the set of all numbers in $(0,1)$ *in whose continued fraction expansion only finitely many integers appear as partial quotients. Then*

$$\mu(S) = 0.$$

Remark 7.2.2. The S has cardinality $2^{\aleph_0}$. Indeed, just letting a_i take the values 1 or 2 gives us a subset of S with cardinality $2^{\aleph_0}$, so $|S| \geq 2^{\aleph_0}$. On the other hand, $S \subset \mathbb{R}$ so $|S| \leq 2^{\aleph_0}$.

Proof. For $M \in \mathbb{N}_{\geq 2}$, let

$$S_M = \{x = [0, a_1, a_2, \dots] : a_i < M, \forall i\}.$$

Since $S_2 \subset S_3 \subset S_4 \subset \dots$, we have $\mu(S) = \lim_{M \to \infty} \mu(S_M)$, and it is thus enough to prove that for all M,

$$\mu(S_M) = 0.$$

For the calculations to follow, it will be convenient to define

$$\gamma = \mu\left(E\left(\begin{smallmatrix}1 & 2 & \dots & n\\ k_1 & k_2 & \dots & k_n\end{smallmatrix}\right)\right), \qquad \tau = 1 - \frac{1}{3M}.$$

Now, suppose that $k_1, \dots, k_n$ are all less than M. For any $k \geq 1$ we have by Lemma 7.1.3:

$$\mu\left(E\left(\begin{smallmatrix}1 & 2 & \dots & n & n+1\\ k_1 & k_2 & \dots & k_n & k\end{smallmatrix}\right)\right) > \frac{\gamma}{3k^2}.$$

This implies that

$$\mu\left(\coprod_{k=M}^{\infty} E\left(\begin{smallmatrix}1 & 2 & \dots & n & n+1\\ k_1 & k_2 & \dots & k_n & k\end{smallmatrix}\right)\right) > \frac{\gamma}{3}\sum_{k=M}^{\infty}\frac{1}{k^2}.$$

We have the estimate

$$\sum_{k=M}^{\infty}\frac{1}{k^2} > \int_M^{\infty}\frac{dt}{t^2} = \frac{1}{M}.$$

We conclude that

$$\mu\left(\coprod_{k<M} E\left(\begin{smallmatrix}1 & 2 & \dots & n & n+1\\ k_1 & k_2 & \dots & k_n & k\end{smallmatrix}\right)\right) = \mu\left(E\left(\begin{smallmatrix}1 & 2 & \dots & n\\ k_1 & k_2 & \dots & k_n\end{smallmatrix}\right)\right) - \mu\left(\coprod_{k=M}^{\infty} E\left(\begin{smallmatrix}1 & 2 & \dots & n & n+1\\ k_1 & k_2 & \dots & k_n & k\end{smallmatrix}\right)\right)$$
$$< \gamma(1 - \frac{1}{3M}) = \gamma\tau.$$

In this estimate $k_1, \dots, k_n$ are still fixed, and γ depends on them. But now, summing this inequality over all $(k_1, k_2, \dots, k_{n+1})$ such that $k_i < M$ for all i, we find

$$\mu\left(\coprod_{(k_1,\dots,k_{n+1})<M} E\left(\begin{smallmatrix}1 & 2 & \dots & n & n+1\\ k_1 & k_2 & \dots & k_n & k_{n+1}\end{smallmatrix}\right)\right)$$
$$< \tau \cdot \mu\left(\coprod_{(k_1,\dots,k_n)<M} E\left(\begin{smallmatrix}1 & 2 & \dots & n\\ k_1 & k_2 & \dots & k_n\end{smallmatrix}\right)\right)$$
$$< \dots < \tau^n \cdot \mu\left(\coprod_{k_1<M} E\left(\begin{smallmatrix}1\\ k_1\end{smallmatrix}\right)\right) < \tau^n.$$

As S_M is the limit of the sets $\coprod_{(k_1,\dots,k_n)<M} E\left(\begin{smallmatrix}1 & 2 & \dots & n\\ k_1 & k_2 & \dots & k_n\end{smallmatrix}\right)$ as $n \to \infty$, we find that

$$\mu(S_M) \leq \lim_{n \to \infty} \tau^n = 0.$$

□

Corollary 7.2.3. *With probability* 1*, a number* x *chosen at random from the interval* $(0, 1)$ *will have the property*

$$\overline{\lim}\, a_i(x) = +\infty.$$

We call a real number θ **badly approximable** if there is a positive constant C such that

$$|\theta - \frac{p}{q}| \geq \frac{C}{q^2},$$

for all p, q integers. One can prove the following theorem.

Theorem 7.2.4. *The set of badly approximable numbers is precisely the set $\bigcup_{n\in\mathbb{Z}} n + S$, where S is as in Theorem 7.2.1.*

7.3. **Some ideas from Ergodic theory.** In his book, Khinchin proves many beautiful theorems of measure-theoretic flavour. The most striking one is due to Khinchin himself and is the following.

Theorem 7.3.1 (Khinchin). *With probability* 1[8]

$$\sqrt[n]{a_1(x)a_2(x)\cdots a_n(x)} \longrightarrow \prod_{r=1}^{\infty}\left(1+\frac{1}{r(r+2)}\right)^{\log_2(r)} \approx 2.685\ldots$$

The infinite product appearing in the theorem is called **Khinchin's constant**. Khinchin's proof is complicated. There is a much better proof that uses ideas from Ergodic theory. We will only sketch it, citing some difficult theorems and leaving some details as exercises.

Consider the operator

$$T : (0,1) \to (0,1), \quad T([0, a_1, a_2, \ldots]) = [0, a_2, a_3, \ldots].$$

Otherwise said

$$T(x) = \frac{1}{x} - \lfloor \frac{1}{x} \rfloor.$$

Define a measure ν on subsets in the Borel σ-algebra $\mathscr{B}([0,1])$ of $[0,1]$:

$$\nu(E) = \frac{1}{\log(2)} \int_E \frac{dx}{1+x}.$$

Like the Lebesgue measure, ν is a regular measure, giving measure 1 to $[0,1]$, and the same argument we sketched for the Lebesgue measure implies that it is determined by its values on open intervals (α, β) that we can easily calculate:

$$\nu((\alpha,\beta)) = \frac{1}{\log(2)} \int_\alpha^\beta \frac{dx}{1+x} = \frac{1}{\log(2)}(\log(1+\beta) - \log(1+\alpha)) = \frac{1}{\log(2)} \log\left(\frac{1+\beta}{1+\alpha}\right).$$

The key point is that T is **measure-preserving**, meaning for every set $E \in \mathscr{B}([0,1])$,

$$\nu(T^{-1}(E)) = \nu(E).$$

This is not very hard to prove. One first proves that $\mu_1(E) := \nu(T^{-1}(E))$ is likewise a regular measure on $(0,1)$, hence determined by its values on open intervals. A bit of thought shows then that it is enough to prove the following claim.

Exercise 7.3.2. ★ For any $0 \leq \beta \leq 1$,

$$\nu(T^{-1}(0,\beta)) = \nu((0,\beta)).$$

[8]Namely the set of $x \in (0,1)$ for which the following statement fails has measure 0.

A deeper fact, yet at the level of the first few classes in a course in ergodic theory, is that the transformation T is **ergodic**. This means that

$$T^{-1}(E) = E \Longrightarrow \nu(E) \in \{0,1\}.$$

For example, real world transformations, such as kneading dough, stirring a coffee cup, are ergodic transformations. Rotating a disc around its centre is not.

We want to use the Ergodic Theorem, which we will not prove here, although, once more, the proof is usually given in any introductory course in ergodic theory.[9] The theorem is much more general than the very special case we consider here.

Theorem 7.3.3 (Birkhoff's Ergodic Theorem)**.** [10] *For any "reasonable" function* f[11] *on* $[0,1]$*, for almost all* $x \in [0,1]$*,*

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} f(T^n x) = \int_0^1 f d\nu := \frac{1}{\log(2)}\int_0^1 \frac{f(x)}{1+x}dx.$$

The sum on the left hand side in the theorem is called the **time average**, since one thinks about T as a transformation of $[0,1]$ where the trajectory of a point x in time is $Tx, T^2x, T^3x, \ldots$ (T^n is the composition of T with itself n-times) and the sum is the average along the trajectory as more and more time passes. The integral appearing in the theorem is called the **space average** as it gives the average value of the function on the space.

We apply the Ergodic Theorem in a variety of situations:

(1) Let

$$f(x) = \log(a_1(x)).$$

This is a function that is piece-wise continuous, even piece-wise constant since $a_1(x)$ is like that. It follows that

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} \log(a_1(T^n x)) = \lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} \log(a_n(x)) = \frac{1}{\log(2)}\int_0^1 \frac{\log(a_1(x))}{1+x}dx. \tag{4}$$

It remains to analyze the right hand side and derive Khinchin's theorem. Recall that

$$a_1(x) = k, \quad \text{for } x \in (\frac{1}{k+1}, \frac{1}{k}).$$

Thus,

$$\begin{aligned}
\frac{1}{\log(2)}\int_0^1 \frac{\log(a_1(x))}{1+x}dx &= \frac{1}{\log(2)}\sum_{k=1}^{\infty} \log(k)\cdot \int_{\frac{1}{k+1}}^{\frac{1}{k}} \frac{1}{1+x}dx \\
&= \frac{1}{\log(2)}\sum_{k=1}^{\infty} \log(k)\cdot(\log(1+1/k) - \log(1+\frac{1}{k+1})) \\
&= \sum_{k=1}^{\infty} \log((1+\frac{1}{k(k+2)})^{\log(k)/\log(2)}) \\
&= \sum_{k=1}^{\infty} \log((1+\frac{1}{k(k+2)})^{\log_2(k)})
\end{aligned}$$

[9] The proof can be found, for example, in P. Halmos, *Lectures on Ergodic Theory*.

[10] There is a subtle point here. A priori the Theorem says that the set of exceptions has measure 0 relative to the measure ν, but it is easy to check that this implies it has measure 0 in the usual (Lebesgue) measure too.

[11] The exact condition is that for all $B \in \mathscr{B}([0,1])$ also $f^{-1}(B) \in \mathscr{B}([0,1])$. This include all functions that are piece-wise continuous, for example.

Substituting in Equation (4), and exponentiating, we find Khinchin's theorem!

(2) Let

$$f(x) = \begin{cases} 1, & a_1(x) = k; \\ 0, & \text{else.} \end{cases}$$

Using the ergodic theorem we can deduce that for almost all $x \in (0,1)$ the frequency of k in the partial quotients of x, namely in the sequence $\{a_i(x)\}_{i=1}^{\infty}$, is

$$\frac{1}{\log(2)} \log\left(\frac{(k+1)^2}{k(k+2)}\right).$$

For example, the frequency of 1 among the $a_i(x)$ is about 41.5% for almost every x. On the other hand, the frequency of 9 is only about 1.4%. The proof is left as an exercise.

(3) Let $f(x) = a_1(x)$ and deduce that with probability 1,

$$\lim_{n \to \infty} \frac{a_1(x) + \cdots + a_n(x)}{n} = \infty.$$

The proof is left as an exercise.

Exercise 7.3.4. ★ What can we deduce by using the function $f(x) = \begin{cases} 1, & a_1(x) \text{ is prime;} \\ 0, & \text{else.} \end{cases}$

Exercise 7.3.5. ★ Give a proof based on the Ergodic Theorem for Theorem 7.2.1.

8. The Hausdorff dimension of some sets defined by continued fractions

We have seen that the set of all real numbers $x \in (0,1)$ whose continued fraction expressions has bounded partial quotients, namely the set

$$S = \{x = [0, a_1, a_2, \dots] : \exists M \text{ such that } a_i \leq M, \forall i\},$$

has measure zero and cardinality $2^{\aleph_0}$. It follows that the set

$$E(\{1,2\}) := \{[0, a_1, a_2, a_3, \dots] : a_i \in \{1,2\}\}$$

also has measure zero and cardinality $2^{\aleph_0}$. It is a natural question to ask whether there are other points of view on subsets of $\mathbb{R}$ that will show that in some sense these sets are big.

We will introduce the notion of Hausdorff dimension of a set A, $\dim_H(A)$, as another perspective on complexity of sets. It will satisfy $\dim_H(\{x\}) = 0$ for any point x and even $\dim_H(\mathbb{Q}) = 0$; it will also satisfy $\dim_H([0,1]) = \dim_H([0,1] \setminus \mathbb{Q}) = 1$. These results just tell us that the definitions are sensible. It starts to get interesting when we find that

$$0 < \dim_H(E(\{1,2\})) < 1, \quad \dim_H(S) = 1,$$

thus providing a new perspective in which the "presence" of these measure 0-sets is nonetheless non-trivial.

In developing the theory we will be forced to be brief – supplying full details would be a course in itself; we will give careful definitions and prove certain simple properties in order to have some sort of intuition as to what's happening and then specialize to sets defined by continued fractions. For a full treatment see Falconer's book, which we follow here.

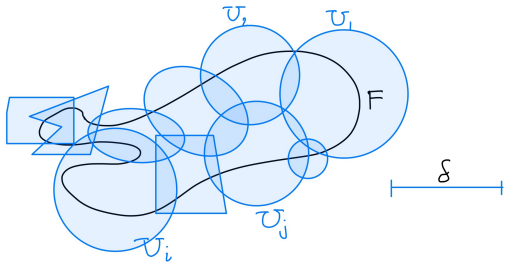
8.1. **δ-covers and the Hausdorff dimension.** Let $U \subset \mathbb{R}^n$. We denote by $|U|$ its **diameter**.

$$|U| = \sup\{\|x - y\| : x, y \in U\}.$$

The notation is not ideal, because $|U|$ was also used to denote the cardinality of U; hopefully, no confusion will occur, and when there is danger of such, we will clarify what the notation means in that particular place.

Let $\delta \in \mathbb{R}_{\geq 0}$. A **$\delta$-cover** of a set F in $\mathbb{R}^n$ is a finite, or countable, collection of sets $\{U_i\}_{i \in \mathscr{I}}$ such that

- $F \subseteq \bigcup_{i \in \mathscr{I}} U_i$
- $|U_i| \leq \delta, \quad \forall i.$



For $s > 0$ define

$$\mathcal{H}^s_\delta(F) = \inf\{\sum_{i \in \mathscr{I}} |U_i|^s : \{U_i\}_{i \in \mathscr{I}} \text{ a } \delta\text{-cover of } F\}.$$

Namely, we are trying to cover F most efficiently using δ-covers and measure this efficiency by the quantity $\sum_{i \in \mathscr{I}} |U_i|^s$. It is not entirely clear what this does... Let us make some observations:

- When $\delta' < \delta$ any δ'-cover of F is also a δ-cover of F. So when calculating $\mathcal{H}^s_\delta(F)$ we are taking the infimum over a larger set of coverings than the set of covering used to calculate $\mathcal{H}^s_{\delta'}(F)$. Hence,

  $$\delta' \leq \delta \Longrightarrow \mathcal{H}^s_{\delta'}(F) \geq \mathcal{H}^s_\delta(F).$$

  Consequently, the following limit, which might well be $+\infty$, always exists:

  $$\mathcal{H}^s(F) := \lim_{\delta \to 0} \mathcal{H}^s_\delta.$$

- Note that for any $\delta < 1$ (and those suffice to calculate $\mathcal{H}^s(F)$), the function $\mathcal{H}^s_\delta(F)$ is non-increasing in s. This is because for any U of diameter at most δ (so smaller than 1) $|U|^s \geq |U|^t$ if $s \leq t$. Thus, comparing cover-by-cover, we deduce that

  $$s \leq t \Longrightarrow \mathcal{H}^s(F) \geq \mathcal{H}^t(F).$$

  We can be more precise: for every $\delta < 1$ cover $\{U_i\}$ and $s \leq t$, we have

  $$\sum_i |U_i|^t = \sum_i |U_i|^{t-s} |U_i|^s \leq \delta^{t-s} \sum_i |U_i|^s.$$

  Thus,

  $$s \leq t \Longrightarrow \mathcal{H}^t(F) \leq \delta^{t-s} \mathcal{H}^s(F), \quad \forall 0 < \delta < 1.$$

The last point allows us to conclude the following proposition.

Proposition 8.1.1. *Let $F \subset \mathbb{R}^n$. If there exists an s such that $\mathcal{H}^s(F) < \infty$ then for all $t > s$, $\mathcal{H}^t(F) = 0$. Thus, the function $s \mapsto \mathcal{H}^s(F)$ must look as in Figure 3, where*

$$s_0 = \sup\{s : \mathcal{H}^s(F) = \infty\} = \inf\{s : \mathcal{H}^s(F) = 0\}.$$

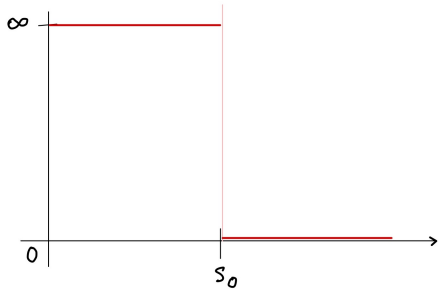


FIGURE 3. The function $s \mapsto \mathcal{H}^s(F)$

The **Hausdorff dimension**[12] of F is, by definition, the value s_0. We denote the Hausdorff dimension $\dim_H(F)$. Thus,

$$\dim_H(F) = \sup\{s : \mathcal{H}^s(F) = \infty\} = \inf\{s : \mathcal{H}^s(F) = 0\}.$$

We remark that the value $\mathcal{H}^{s_0}(F)$ may be finite (including 0), or infinite. However, we can conclude the following.

Corollary 8.1.2. *If there is some s such that $0 < \mathcal{H}^s(F) < \infty$, then $s = \dim_H(F)$.*

Example 8.1.3. To get convinced that this might be a good definition, let us look at the interval $[0,1]$ and take $s = 1$. For any δ-cover $\{U_i\}$ let $a_i = |U_i|$. The estimate

$$\sum_i a_i \geq 1$$

holds, because $a_i \geq \mu(U_i)$ (Lebesgue measure) and

$$1 \leq \mu(\cup_i U_i) \leq \sum_i \mu(U_i) \leq \sum_i a_i.$$

This holds for any cover and any δ. Further, we can certainly find such covers with $\sum a_i = 1$; indeed, dividing $[0,1]$ into disjoint intervals will do the trick. It follows that $\mathcal{H}^1([0,1]) = 1$ and, by Corollary 8.1.2, $\dim_H([0,1]) = 1$. Thus, for "ordinary" sets the Hausdorff dimension returns the expected value. The interesting feature, though, is that for "non-ordinary" sets it may return a fractional value, as we shall see.

The following theorem lists some basic properties of the Hausdorff dimension. We leave the first four claims as an exercise in unravelling the definitions. We highly recommend doing them.

Theorem 8.1.4. *The Hausdorff dimension has the following properties for subsets of $\mathbb{R}^n$:*

(1) If $E \subset F$ then $\dim_H(E) \leq \dim_H(F)$.
(2) $\dim_H(\cup_{i=1}^{\infty} F_i) = \sup_i\{\dim_H(F_i)\}$.
(3) If F is countable, $F = \{x_1, x_2, \dots\}$, then $\dim_H(F) = 0$.
(4) If $f\colon \mathbb{R}^n \to \mathbb{R}^n$ is bi-Lipschitz (namely, there exist positive real constants c_1, c_2 such that

$$c_1\|x-y\| \leq \|f(x) - f(y)\| \leq c_2\|x-y\|, \quad \forall x, y \in \mathbb{R}^n)$$

then,

$$\dim_H(F) = \dim_H(f(F)).$$

(5) If $F \subseteq \mathbb{R}^n$ contains a non-empty open set then $\dim_H(F) = n$.
(6) If $F \subseteq \mathbb{R}^n$ is an m-dimensional smooth manifold, $\dim_H(F) = m$.

[12]The Hausdorff dimension is sometimes called the **Hausdorff-Besicovitch dimension**.

Exercise 8.1.5. Prove that for every s, $\mathcal{H}^s(\cup_{i=1}^{\infty} F_i) \leq \sum_{i=1}^{\infty} \mathcal{H}^s(F_i)$.

Exercise 8.1.6. Prove parts (1) - (4) of Theorem 8.1.4.

Property (4) is very useful in boot-strapping results. Consider for example, a bi-Lipschitz function $f\colon \mathbb{R} \to \mathbb{R}$ and the function $\phi\colon \mathbb{R}^2 \to \mathbb{R}^2$, $\phi(x,y) = (x, f(x)+y)$, taking the unit interval $A = [0,1] = \{(x,0) : 0 \leq x \leq 1\}$ to the set $\phi(A) = \{(x, f(x))\}$, which is just the graph of f over the interval $[0,1]$.

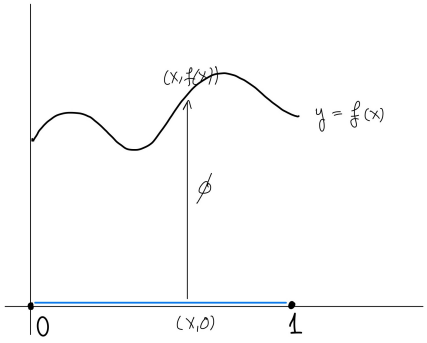
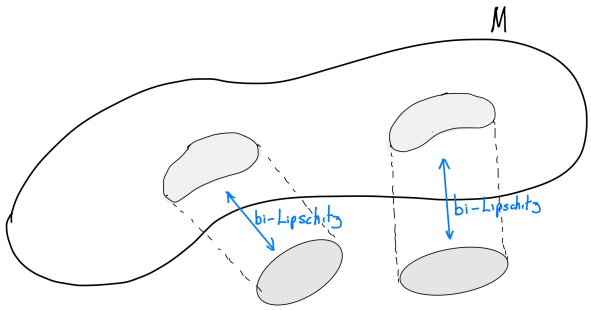


FIGURE 4. The function ϕ

This function ϕ is also bi-Lipschitz and so $\dim_H(\phi(A)) = \dim_H(A)$. We proved before that $\dim_H(A) = 1$ when A is viewed as a subset of $\mathbb{R}$. It requires some thought, but one can show that $\dim_H(A) = 1$ also when A is viewed as a subset of $\mathbb{R}^2$. So we may conclude that the graph of ϕ is also of dimension 1.

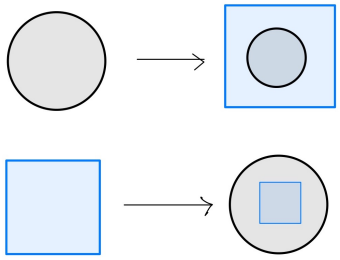
Property (6) requires some work to prove, but the idea is very similar to the example we have just considered. Suppose we already know that $\dim([a,b]^m) = m$ for $a < b$ (which follows from (5)), whether this cube is considered in $\mathbb{R}^m$ or in $\mathbb{R}^n$. Then, the idea is that by the implicit function theorem, any smooth manifold can be exhibited locally as the graph of a C^∞-function over such cubes (that appear in the diagram as ovals; mea culpa).



Such functions are bi-Lipschitz so, at least locally, a smooth m-dimensional manifold has Hausdorff dimension m. Using property (2) we can get the full statement in (6).

Let's look at claim (5) and try and be a little bit "more honest" about its proof.

- First, all closed cubes $\prod_{i=1}^{n}[a_i, b_i]$ with non-empty interior, i.e. with $a_i < b_i, \forall i$, have the same dimension as they are all related by bi-Lipschitz maps. We remark that this dimension is also equal to the dimension of any closed ball $B[a,r]$ of positive radius, because, again they are related by bi-Lipschitz maps, and we also use property (1).



- As F contains a closed ball with a non-empty interior, and is contained in a countable union of such balls (e.g. $\cup_{N=1}^{\infty} B[0,N]$), using (2) we conclude that $\dim_H(F) = \dim_H(B[0,1]) = \dim_H([0,1]^n)$.
- Show $\dim_H(([0,1]^n) \leq n$. It is enough to show that
$$\forall n_1 > n, \mathcal{H}^{n_1}([0,1]^n) = 0.$$
For this, it is enough to show that for every $\epsilon > 0$, for all $\delta > 0$, there is a δ-cover $\{U_i\}$ such that
$$\sum_i |U_i|^{n_1} < \epsilon.$$
Indeed, this implies that $\mathcal{H}^{n_1}_{\delta}(([0,1]^n) < \epsilon$ for all $\delta > 0$, which implies that $\mathcal{H}^{n_1}(([0,1]^n) < \epsilon$ for all $\epsilon > 0$. This implies that $\mathcal{H}^{n_1}(([0,1]^n) = 0$.

  In fact, it is enough to show that for every $\epsilon > 0$ for all $N \gg 0$ there is a $\frac{1}{N}$-cover $\{U_i\}$ such that $\sum_i |U_i|^{n_1} < \epsilon$.

  Divide the cube $[0,1]^n$ into N^n cubes that are shifts of the cube $[0, \frac{1}{N}]^n$. The radius of a cube $[0,r]^n$ is $\|(r,r,\ldots,r)\| = r\sqrt{n}$.

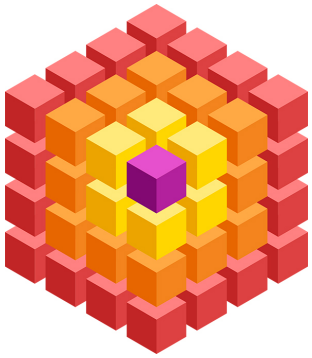


FIGURE 5. Partitioning the cube (The graphics is taken from Deke McClelland's blog)

For this cover,
$$\sum_i |U_i|^{n_1} = N^n(\tfrac{\sqrt{n}}{N})^{n_1} = C \cdot N^{n-n_1} \xrightarrow{N\to\infty} 0$$

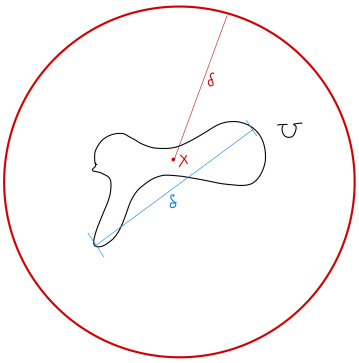
- Show that $\mathcal{H}^n([0,1]^n) > 0$. We use a similar argument as we have used for the interval $[0,1]$. Namely, the existence of a Lebesgue measure μ on $\mathbb{R}^n$. We didn't discuss this before, but this is similar to the Lebesgue measure on $\mathbb{R}$ as in §3.2. Namely, μ is a again a function on the Borel σ-algebra $\mathscr{B}(\mathbb{R}^n)$ of $\mathbb{R}^n$, which is the minimal collection of subsets of $\mathbb{R}^n$ that contains every open set and is closed under complements, countable unions and countable intersections:
$$\mu\colon \mathscr{B}(\mathbb{R}^n) \to \mathbb{R}_{\geq 0} \cup \{\infty\}.$$

It is again a regular measure such that (1) $\mu(\coprod_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$, and (2) $\mu(\prod_{i=1}^{n}[a_i, b_i]) = \prod_{i=1}^{n}(b_i - a_i)$.

For every δ-cover[13] $\{U_i\}$ we have

$$1 = \mu([0,1]^n) \leq \sum_i \mu(U_i) \leq C \sum_i |U_i|^n,$$

where C is a constant depending only on n and the last inequality comes from the fact that every set U_i of diameter δ is contained in a ball of radius δ, say.



We conclude that

$$\mathcal{H}^n([0,1]^n) \geq \frac{1}{C} > 0.$$

Remark 8.1.7. In fact, one can prove that $\mathcal{H}^n([0,1]^n) = \frac{2^n}{\omega_n}$, where ω_n is the volume of the unit ball in $\mathbb{R}^n$. More generally, one can prove that *if $F \in \mathscr{B}(\mathbb{R}^n)$ is a Borel set, then*

$$\mathcal{H}^n(F) = \frac{\mu(F)}{\omega_n}.$$

We have

$$\omega_n = \begin{cases} \frac{\pi^k}{k!}, & n = 2k \text{ even;} \\ \frac{2^{2k+1}k!\pi^k}{(2k+1)!} & n = 2k+1 \text{ odd.} \end{cases} \tag{5}$$

8.2. **The dimension of the Cantor set.** So far we haven't really seen that the Hausdorff dimension provides us with anything new. We will now calculate the dimension of the **Cantor set** $\mathscr{C}$ in $[0,1]$ and see that it is strictly between 0 and 1.

Base 10 expansion, i.e., decimal expansions, are the description of any real number r as

$$r = a_n \cdots a_1 a_0 . a_{-1} a_{-2} \cdots = a_n \cdot 10^n + \cdots + a_1 \cdot 10 + a_0 + a_{-1} \cdot 10^{-1} + a_{-2} \cdot 10^{-2} + \ldots.$$

Or, more succinctly,

$$r = \sum_{s=-n}^{\infty} a_{-s} 10^{-s},$$

where the $a_i \in \{0, \ldots, 9\}$. The expansion is essentially unique; the only ambiguity comes from tails of the form $\cdots a_{i-1} a_i 99999 \cdots = \cdots a_{i-1}(a_i + 1)00000 \ldots$, if $a_i \neq 9$. Let us agree to prefer the latter.

[13]There is a delicate point we are sweeping under the rug. Namely, that when we define the notion of Hausdorff dimension we can consider only δ-covers by sets U_i that are in $\mathscr{B}(\mathbb{R}^n)$. This is true, and not hard to prove. Indeed, we can replace any U_i that appears with its closure.

The same can be done to any natural basis. Let $N \geq 2$ be an integer. Then any real number has a base N expansion

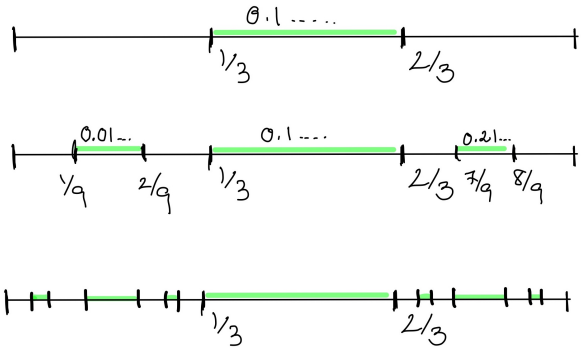
$$r = a_n \cdots a_1 a_0 . a_{-1} a_{-2} \cdots = a_n N^n + \cdots + a_1 N + a_0 + a_{-1} N^{-1} + a_{-2} N^{-2} + \cdots = \sum_{s=-n}^{\infty} a_{-s} N^{-s},$$

where the $a_i \in \{0, \ldots, N-1\}$. Again, the expansion is essentially unique and the only ambiguity comes from tails of the form $\cdots a_{i-1} a_i x x x x x x \cdots = \cdots a_{i-1}(a_i + 1)00000 \ldots$, if $a_i \neq N-1$ and $x = N-1$. (For example, for $N = 2$, $0.11111111 \cdots = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots = 1$, and for $N = 3$ we have $0.22222222 \cdots = \frac{2}{3} + \frac{2}{9} + \frac{2}{27} + \cdots = 1$.) Let us agree to prefer the former just for describing the Cantor set.

With these conventions the Cantor set is defined as

$$\mathscr{C} = \{x = 0.x_1 x_2 x_3 \ldots \text{ (base 3 expansion)} | x_i \in \{0, 2\}, \forall i\} = \left\{ \sum_{i=1}^{\infty} \frac{x_i}{3^i} : x_i \in \{0, 2\}, \forall i \right\}.$$

It is a closed set in $[0, 1]$ which is what remains after repeatedly *removing* the middle open thirds (in green) of intervals.



By calculating the measure of the complement of the Cantor set we find that $\mathscr{C}$ has measure 0, but cardinality equal to that of the real numbers:

$$\mu(\mathscr{C}) = 0, \quad |\mathscr{C}| = 2^{\aleph_0}.$$

Let assume for now that for $d = \dim_H(\mathscr{C})$ we actually have $0 < \mathcal{H}^d(C) < \infty$ (which is true). The idea of the following calculation is that of an iterated function system, which we will soon develop in detail; it makes use of the self-similarity features of the Cantor set.

Let

$$s_1(x) = \frac{x}{3}, \qquad s_2(x) = \frac{x}{3} + \frac{2}{3}.$$

In terms of base 3-expansion, we can express these transformation by

$$s_1(0.x_1 x_2 x_3 \ldots) = 0.0 x_1 x_2 x_3 \ldots, \qquad s_2(0.x_1 x_2 x_3 \ldots) = 0.2 x_1 x_2 x_3 \ldots.$$

It is now clear that

$$\mathscr{C} = s_1(\mathscr{C}) \coprod s_2(\mathscr{C}).$$

Exercise 8.2.1. Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a function such that for some $\alpha > 0$

$$\|f(x) - f(y)\| = \alpha \|x - y\|.$$

Then,

$$\mathcal{H}^s(f(F)) = \alpha^s \mathcal{H}^s(F).$$

For the functions s_1, s_2 we have $\alpha = 1/3$ and we apply the exercise to the Cantor set $\mathscr{C}$ with $s = d = \dim_H(\mathscr{C})$. Note that for δ small enough we may assume in calculating $H^s_\delta(\mathscr{C})$ that we only consider δ-covers of $s_1(\mathscr{C}) \coprod s_2(\mathscr{C})$ that are a disjoint union of a δ-cover of $s^1(\mathscr{C})$ and a δ-cover of $s^2(\mathscr{C})$. And vice-versa, δ-covers of $s^i(\mathscr{C})$ that are disjoint with each other induce a δ-cover of $\mathscr{C}$. Hence,

$$\begin{aligned}\mathcal{H}^d(\mathscr{C}) &= \mathcal{H}^d(s_1(\mathscr{C}) \coprod s_2(\mathscr{C})) \\ &= \mathcal{H}^d(s_1(\mathscr{C})) + \mathcal{H}^d(s_2(\mathscr{C})) \\ &= 2 \cdot \left(\frac{1}{3}\right)^d \mathcal{H}^d(\mathscr{C}).\end{aligned}$$

Thus, $3^d = 2$ and we conclude that

$$\dim_H(\mathscr{C}) = \frac{\log(2)}{\log(3)} \approx 0.6309\ldots$$

This result demonstrates that although $\mu(\mathscr{C}) = 0$, the Cantor set $\mathscr{C}$ still has "a non-zero presence".

Exercise 8.2.2. Let A be the set of all numbers in $[0,1]$ whose base 5 expansion only contains the digits $0, 2$ and 4. Let $d = \dim_H(A)$ and assume again that $0 < \mathcal{H}^d(A) < \infty$. Calculate $\dim_H(A)$.

Exercise 8.2.3. ★ Let $N \geq 3$ be an odd integer. Let A_N be the set of all numbers in $[0,1]$ whose base N expansion only contains the digits $0, 2, \ldots N-1$. Let $d = \dim_H(A_N)$ and assume again that $0 < \mathcal{H}^d(A_N) < \infty$. Calculate $\lim_{N\to\infty} \dim_H(A_N)$.

The following result, which we give as a guided exercise, gives some information on sets of dimension less than 1. Recall that if $F \subset \mathbb{R}^n$ is a subset, an open set of F is, by definition, the intersection of some open subset of $\mathbb{R}^n$ with F. A set F in $\mathbb{R}^n$ is called **totally disconnected** if for every $x \neq y$ points of F, there exist open disjoint sets U_x, U_y of F, such that $x \in U_x, y \in U_y$ and $F = U_x \cup U_y$.

For example, $\mathbb{Q}$ is totally disconnected, because given two rational numbers $x < y$ choose an irrational ϵ such that $x < \epsilon < y$ then

$$\mathbb{Q} = (\mathbb{Q} \cap (-\infty, \epsilon)) \cup (\mathbb{Q} \cap (\epsilon, \infty)),$$

and

$$x \in U_x := \mathbb{Q} \cap (-\infty, \epsilon), \quad y \in U_y := \mathbb{Q} \cap (\epsilon, \infty).$$

Exercise 8.2.4. Let $F \subseteq \mathbb{R}^n$ be a subset such that $\dim_H(F) < 1$. Prove that F is totally disconnected. Here is a suggestion. Suppose $x \neq y$ are points in F:

- Define
$$f\colon \mathbb{R}^n \to \mathbb{R}, \quad f(s) = \|x - s\|.$$
Prove that this function has the property
$$|f(s) - f(t)| \leq \|s - t\|.$$
- Prove that $\dim_H(f(F)) \leq \dim_H(F)$.
- Let $Z = \mathbb{R} \setminus f(F)$. Prove that Z is dense in $\mathbb{R}$.
- Prove that there is a $z \in Z$ lying between $f(x)$ and $f(y)$ and so the sets $(-\infty, z), (z, \infty)$ separate $f(x)$ and $f(y)$.
- Complete the proof.

Note that the converse does not hold: $\mathbb{R} \setminus \mathbb{Q}$ is totally disconnected but has dimension 1.

8.3. **Iterated function systems.** An iterated function system is a method to construct fractal sets in $\mathbb{R}^n$ and, at the same time, to estimate their dimensions. Many familiar sets, such as the Cantor set, the von Koch snowflake and the Sierpinski cube are examples. More importantly for us, sets whose elements are described by their continued fraction expansions are often constructed this way too.

Let $D \subseteq \mathbb{R}^n$ be a closed set ($D = \mathbb{R}^n$ is allowed). A function

$$s\colon D \to D$$

is called a **contraction** if there is a constant c, $0 < c < 1$, such that

$$\|s(x) - s(y)\| \leq c\|x - y\|, \quad x, y \in D.$$

We note that a contraction is automatically a continuous function.

An **iterated function system** (**IFS**) is a finite set of maps $\{s_1, \dots, s_m\}$, $m \geq 2$, where every s_i is a contraction on D. A non-empty closed subset $F \subseteq D$ is called an **attractor** for the IFS if

$$F = \bigcup_{i=1}^{m} s_i(F).$$

Example 8.3.1. Let $D = [0, 1]$ and let $s_i\colon D \to D$ be the functions

$$s_1(x) = \frac{x}{3}, \quad s_2(x) = \frac{x}{3} + \frac{2}{3}.$$

Then $\{s_1, s_2\}$ is an IFS and, as we have seen before, the Cantor set $\mathscr{C}$ is an attractor.

Example 8.3.2. Let

$$E = \{[a_0, a_1, a_2, \dots] : a_i \in \{1, 2\}\},$$

be the subset of all real numbers in the interval $[1, 3]$ whose continued fraction expansion has partial quotients that are all either 1 or 2. Let D be the interval

$$D = [\frac{1 + \sqrt{3}}{2}, 1 + \sqrt{3}] \approx [1.366\dots, 2.732\dots].$$

Consider the two functions on $\mathbb{R}_{>0}$

$$s_1(x) = 1 + \frac{1}{x}, \quad s_2(x) = 2 + \frac{1}{x}.$$

We claim that $\{s_1, s_2\}$ are an IFS on D. For $x \in D$,

$$|s_i'(x)| = \frac{1}{x^2} \in \left[\frac{1}{(1 + \sqrt{3})^2}, \frac{4}{(1 + \sqrt{3})^2}\right] \subseteq [0.13, 0.54].$$

By the mean-value theorem

$$\left|\frac{s_i(x) - s_i(y)}{x - y}\right| = \left|s_i'(\xi)\right|,$$

and so both s_i are contractions on D. Now, both s_i have a fixed point ζ_i in D:

$$s_1(\frac{1 + \sqrt{5}}{2}) = \frac{1 + \sqrt{5}}{2} \approx 1.618\dots, \qquad s_2(1 + \sqrt{2}) = 1 + \sqrt{2} \approx 2.414\dots$$

This should not surprise us. In terms of continued fractions we have

$$s_1([a_0, a_1, a_2, \dots]) = [1, a_0, a_1, a_2, \dots], \qquad s_2([a_0, a_1, a_2, \dots]) = [2, a_0, a_1, a_2, \dots], \tag{6}$$

and we have seen the values $\frac{1+\sqrt{5}}{2} = [1, 1, 1, \dots]$, $1 + \sqrt{2} = [2, 2, 2, \dots]$ before. Because the fixed point ζ_i of s_i lies in D and s_i is contracting on D, for any y in D we have $|s_i(y) - \zeta_i| \leq |y - \zeta_i|$, it follows that $s_i(D) \subseteq D$.

Finally, we claim that E is an attractor for this IFS. First, the minimal element of E is

$$[1,2,1,2,\dots] = \frac{1+\sqrt{3}}{2},$$

and the maximal element is

$$[2,1,2,1,\dots] = 1+\sqrt{3}.$$

So $E \subset D$. And equation (6) proves that

$$E = s_1(E) \cup s_2(E).$$

8.4. **Attractors of IFS.** We will show that every IFS on a compact set $D \subseteq \mathbb{R}^n$ has an attractor. The assumption that D is compact is only made to simplify the proof; it holds for the applications we have in mind.

Theorem 8.4.1. *Let D be a non-empty compact subset of $\mathbb{R}^n$ and let $\mathscr{I} = \{s_1, \dots, s_m\}, m \geq 2$ be an IFS on D. There is a unique attractor E for $\mathscr{I}$. In fact, define for every k-tuple of integers $(i_1, \dots, i_k)$, such that $1 \leq i_j \leq m$,*

$$S_{(i_1,\dots,i_k)}(D) = (s_{i_1} \circ \cdots \circ s_{i_k})(D),$$

and

$$S^k(D) = \bigcup_{(i_1,\dots,i_k)} S_{(i_1,\dots,i_k)}(D).$$

Let $S^0(D) = D$. Then,

$$E = \bigcap_{k=0}^{\infty} S^k(D).$$

Before the proof, we remark that with some abuse of notation, treating S^1 as a "sum" of maps without defining this formally, we have the following relation

$$S^k(D) = S^1 \circ \cdots \circ S^1(D).$$

Proof. We first argue that

$$S^k(D) \supseteq S^{k+1}(D).$$

Indeed,

$$S^{k+1}(D) = \underset{(k-\text{times})}{S^1 \circ \cdots \circ S^1}(S^1(D)) = S^k(S^1(D)).$$

So we only need to check that $S^1(D) \subseteq D$, and, in fact, just that $s_i(D) \subset D$, which is true by definition.

Now, as each s_i is a contraction, it is continuous and hence so is $s_{(i_1,\dots,i_k)}$. Therefore, $s_{(i_1,\dots,i_k)}(D)$ is compact. Since $S^k(D)$ is a finite union of non-empty compact sets, it is compact too and non-empty. The intersection of a decreasing sequence of the non-empty compact sets, $D \supseteq S^1(D) \supseteq S^2(D) \supseteq \dots$ is also non-empty. Furthermore, since $S^1(S^k(D)) = S^{k+1}(D)$ we have that $S^1(E) = E$, as we wanted.

We leave the unicity as an exercise. (See below.) □

Exercise 8.4.2. Let D be a compact set and let $\mathcal{S}$ be the collection of non-empty compact subsets of D. We will make S into a metric space; the metric is known as the **Hausdorff metric**.

Let $\delta \geq 0$. Define the **δ-neighbourhood** of a set $A \in \mathcal{S}$, denoted A_δ, to be the set

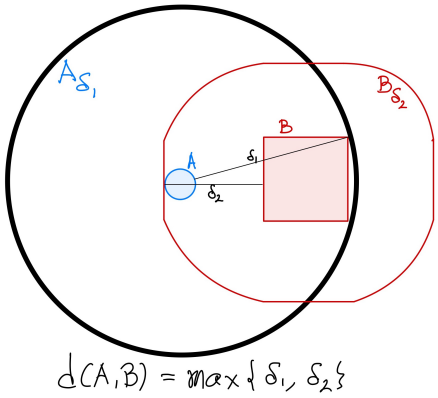
$$A_\delta = \{x \in D : \exists a \in A, \|x - a\| \leq \delta\}.$$

Using that for a fixed $x \in D$, $\inf_{a \in A}\{\|x-a\|\}$ is achieved for some $a_x \in A$, by compactness of A, it is not hard to prove that A_δ is a closed subset of D, hence belongs to $\mathcal{S}$ itself.

Now, given $A, B \in \mathcal{S}$, define

$$d(A,B) = \inf\{\delta : A \subseteq B_\delta \text{ and } B \subseteq A_\delta\}.$$

Prove that d is indeed a metric on $\mathcal{S}$. Namely, that (i) $d(A,B) \geq 0$ with equality iff $A = B$; (ii)



$d(A,B) = d(B,A)$; (iii) $d(A,C) \leq d(A,B) + d(B,C)$.

Exercise 8.4.3. Prove that for $A_i, B_i \in \mathcal{S}$,

$$d(\cup_{i=1}^m A_i, \cup_{i=1}^m B_i) \leq \max_{1 \leq i \leq m} d(A_i, B_i).$$

Exercise 8.4.4. Prove that E is unique in the following sense. If $F \subseteq D$ is a compact non-empty subset of D such that $S^1(F) = F$ then $F = E$. (Consider $d(E,F)$ and apply the previous exercises for $A_i = s_i(E), B_i = s_i(F)$.)

8.5. **Hausdorff dimension of attractors.** In this section we state the main theorem about the dimension, or estimate for the dimension, of the attractor of an IFS. The proof is not very hard,[14] but we have to keep our goal in sight, which is to focus on number theory and not on fractal theory!

We say that an IFS $\{s_1, \ldots, s_m\}, m \geq 2$ on a non-empty compact set D in $\mathbb{R}^n$ satisfies the **open set condition**, if there exists a non-empty open set V of $\mathbb{R}^n$ such that:

(1) $V \subseteq D$,
(2) $s_1(V), \ldots, s_m(V)$ are disjoint, and
(3) $V \supseteq \bigcup_{i=1}^m s_i(V)$.

Theorem 8.5.1. *Let $\{s_1, \ldots, s_m\}, m \geq 2$, be an IFS on a compact set D in $\mathbb{R}^n$ satisfying the open set condition. Let E be the unique compact attractor this system.*

(1) Assume that for every i there exists c_i, such that $0 < c_i < 1$ and

$$\|s_i(x) - s_i(y)\| = c_i \|x - y\|, \forall x, y \in D.$$

Then, $\dim_H(E) = s$, where s is the solution to the equation

$$\sum_{i=1}^m c_i^s = 1.$$

Furthermore, we have $0 < \mathcal{H}^{\dim_H(E)}(E) < \infty$.[15]

[14] The proof can be found in K. Falconer, *Fractal Geometry, Mathematical Foundations and Applications*.

[15] This justifies the computations we did, for example, for Cantor sets.

(2) *If we only have*

$$\|s_i(x) - s_i(y)\| \leq c_i\|x - y\|, \forall x, y \in D,$$

then $\dim_H(E) \leq s$, *where* s *is as above.*

(3) *Suppose that*

$$\|s_i(x) - s_i(y)\| \geq b_i\|x - y\|, \forall x, y \in D,$$

for some $0 < b_i < 1$ *and that*

$$E = \coprod_{i=1}^{m} s_i(E).$$

Then $\dim_H(E) \geq t$, *where* t *is the solution to the equation*

$$\sum_{i=1}^{m} b_i^t = 1.$$

As said, we will not prove the theorem here - the proof can be found in Falconer's book - but only remark that the main idea is rather similar to our calculation of the dimension of the Cantor set $\mathscr{C}$. And, in fact, we will redo this example soon.

Example 8.5.2. The dimension of the unit cube $[0,1]^n$. Let

$$D = [0,1]^n.$$

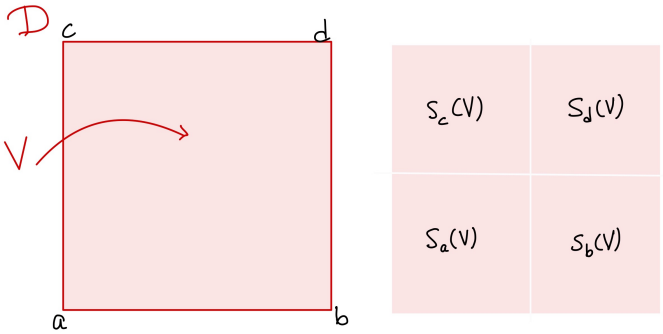
For every vertex v of the cube, define that function

$$s_v \colon D \to D, \quad s_v(x) = \frac{1}{2}x + \frac{1}{2}v.$$

Each s_v is a contraction on D and, in fact,

$$\|s_v(x) - s_v(y)\| = \frac{1}{2}\|x - y\|.$$

This IFS satisfies the open set condition with the open set $V = (0,1)^n$. We also have $D = \cup_v s_v(D)$



and so $D = E$ is the attractor. We find that $\dim_H([0,1]^n) = s$, where s solves the equation $2^n \cdot \left(\frac{1}{2}\right)^s = 1$. Namely, $s = n$, as expected.

Example 8.5.3. The dimension of the Cantor set $\mathscr{C}$. Let

$$D = [0,1], \quad s_1(x) = \frac{x}{3}, \quad s_2(x) = \frac{x}{3} + \frac{2}{3}.$$

This is an iterated function system and $\|s_i(x) - s_i(y)\| = \frac{1}{3}\|x - y\|$. The open set condition holds with $V = (0,1)$. As we have seen, the attractor of this IFS is the Cantor set. Its dimension is the real number s that solves the equation $2 \cdot \left(\frac{1}{3}\right)^s = 1$. Namely,

$$s = \log(2)/\log(3) = 0.6309\ldots.$$

Example 8.5.4. The dimension of the Sierpinski triangle $\widetilde{\triangle}$. Let D be an equilateral triangle in the plane, say with vertices $a_1 = (0,0), a_2 = (1,0), a_3 = (\frac{1}{2}, \frac{\sqrt{3}}{2})$. Let

$$s_i \colon D \to D, \quad s_i(x) = \frac{1}{2}x + \frac{1}{2}a_i, \quad i = 1,2,3.$$

The open set condition holds with the set V being the interior of D. The attractor E of this IFS is, by definition, the **Sierpinski triangle**.

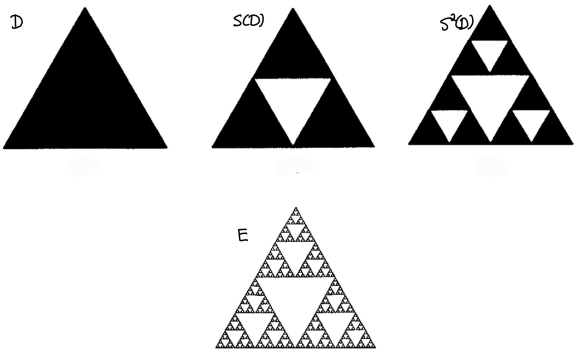


FIGURE 6. The Sierpinski triangle $\widetilde{\triangle} = E$.

A similar calculation gives us

$$\dim_H(\widetilde{\triangle}) = \log(3)/\log(2) = 1.5849\ldots.$$

Example 8.5.5. The dimension of the von Koch snowflake $\mathscr{K}$. We leave some details here to the reader. The iterated function system S consists here of 4 transformations that are each of the form

$$x \mapsto \rho_i(\frac{x}{3}) + \sigma_i,$$

where ρ_i is a certain rotation and σ_i a certain translation that will hopefully be clear from figure. In this case it is easier to construct the attractor as a union of sets, starting from F and applying S repeatedly, we have

$$E = \lim_{k \to \infty} S^k(F).$$

One has to justify all these considerations, including the notion of the limit, but we will leave it to the reader to ponder. The limit, for example, could be in the sense of the Hausdorff metric on the compact sets of D, where D is a circle of radius 3, and in this case, F is drawn from the origin to the point $(3,0)$.

The same type of calculations give us that

$$\dim_H(\mathscr{K}) = \log(4)/\log(3) = 1.2618\ldots.$$

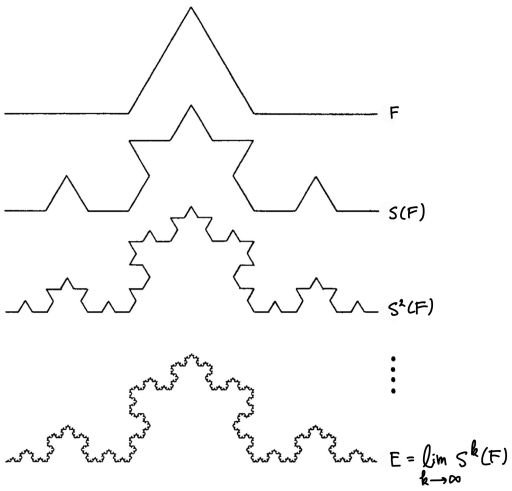


FIGURE 7. The von Koch snowflake $\mathscr{K} = E$.

B. Mandelbrot, who coined the term fractal, and popularized the notion not just in mathematics, but in culture at large, also coined the phrase "how long is the coast of Britain?" [16] The idea being that pictures such as the von Koch curve give a better sense of a coast line than the usual piece-wise smooth lines we tend to use in maps and diagrams. In this context, it is easy to prove that the length of the von Koch curve is infinite and a better measure of its complexity is the fact that its dimension is strictly bigger than 1. To quote Mandelbrot (loc. cit.): "Quantities other than length are thus needed to discriminate between various degrees of complication for a geographical curve."

Exercise 8.5.6. **The dimension of the Sierpinski cube.** Prove that the dimension of the Sierpinski cube is $\log(20)/\log(3) = 2.7268\ldots$. (Picture from Wikipedia commons.)

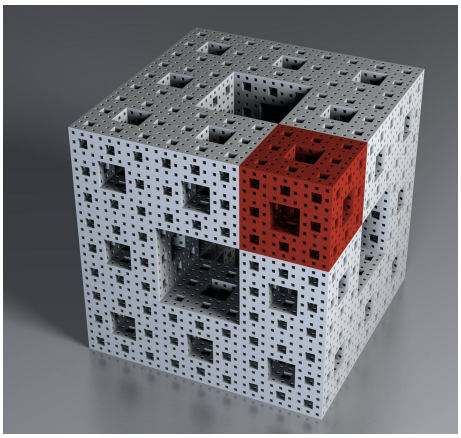


FIGURE 8. The Sierpinski cube.

[16] B. Mandelbrot, How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension, *Science* 156 (3775): 636–638.

Exercise 8.5.7. Let $0 < \alpha < 1$. Consider a set $\mathscr{C}_\alpha$ which is very similar to the Cantor set. At each step we remove an interval of length 2α which is located symmetrically. Thus, the case of the Cantor set itself is when $\alpha = 1/3$.

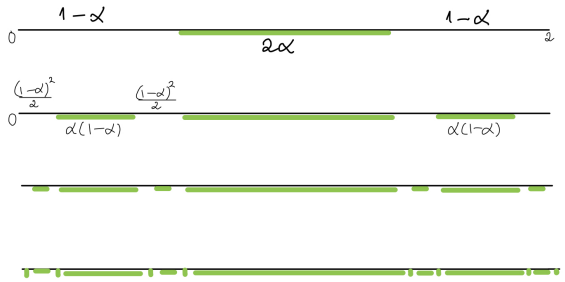


FIGURE 9. Generalized Cantor set $\mathscr{C}_\alpha$.

Calculate the cardinality, the measure and the dimension of the set $\mathscr{C}_\alpha$.

8.6. **Hausdorff dimension of sets defined by continued fractions.** We return now to continued fractions. Let $m \geq 2$ be an integer and let

$$E(\{1,2,\dots,m\}) = \{[a_0,a_1,a_2,\dots] : 1 \leq a_i \leq m, \forall i\}.$$

This set is very similar to the sets S_m we considered in Theorem 7.2.1. In fact, $E(\{1,2,\dots,m\}) = \coprod_{i=1}^m i + S_{m+1}$ and so it is clear that

$$\mu(E) = 0, \quad |E| = 2^{\aleph_0}.$$

We would like to calculate $\dim_H(E(\{1,2,\dots,m\}))$. It turns out to be difficult and so we will only approximate this dimension based on theorems we have already mentioned.

To begin with, the minimal element of E is

$$\min_{x\in E(\{1,2,\dots,m\})}(x) = [1,m,1,m,\dots] =: \alpha_m = \frac{m+\sqrt{m^2+4m}}{2m},$$

and the maximal element is

$$\max_{x\in E(\{1,2,\dots,m\})}(x) = [m,1,m,1,\dots] =: A_m = \frac{m+\sqrt{m^2+4m}}{2}.$$

Let

$$D_m = [\alpha_m, A_m].$$

Let

$$s_i(x) = i + \frac{1}{x}, \quad i = 1,2,\dots,m,$$

or, in terms of continued fractions,

$$s_i([a_0,a_1,a_2,\dots]) = [i,a_0,a_1,a_2,\dots].$$

We show that the s_i are an IFS on D_m; $E(\{1,2,\dots,m\})$ is the attractor of this system. In fact, this requires justification, but not to distract from the main point, we will address it once we have completed the calculation.

- We first show that s_a has a fixed point on D_m. Consider the equation $f_a = a + 1/f_a$. Such an f_a is a fixed point of s_a. We calculate that
$$f_a = \frac{a + \sqrt{a^2+4}}{2} \in D_m.$$
In terms of continued fractions, $f_a = [a, a, a, \dots] \in E \subset D_m$.
- Secondly, we show that each s_a is contracting on D_m. Indeed, since
$$s_a'(x) = -1/x^2,$$
we have
$$0 < A_m^{-2} = \inf_{x \in D_m} \{|s_a'(x)|\}, \quad \alpha_m^{-2} = \sup_{x \in D_m} \{|s_a'(x)|\} < 1,$$
and so, by the mean-value theorem,
$$\forall x, y \in D_m, \quad A_m^{-2}|x-y| \leq |s_a(x) - s_a(y)| \leq \alpha_m^{-2}|x-y|.$$

It follows that each s_i preserves D_m and so $\{s_1, \dots, s_m\}$ is an IFS on D_m.

We may apply Theorem 8.5.1 and conclude
$$\gamma_m \leq \dim_H(E(\{1, 2, \dots, m\})) \leq \Gamma_m,$$
where γ_m and Γ_m are the solutions to the equations
$$mA_m^{-2\gamma_m} = 1, \qquad m\alpha_m^{-2\Gamma_m} = 1.$$
Namely,
$$\frac{\log(m)}{2\log(A_m)} \leq \dim_H(E(\{1, 2, \dots, m\})) \leq \frac{\log(m)}{2\log(\alpha_m)}. \tag{7}$$
The upper bound is uninteresting, it is greater than 1 for all m. However, the method can be improved to provide better lower bounds and non-trivial upper bounds.

The dimension of $E(\{1,2\})$ in particular gained a lot of attention. The world-record for calculating its dimension seems to be

$$\dim_H(E(\{1,2\})) = 0.531280506277205141624468647368471785493059$$
$$1090183987798883978039275295356438313459181095701811852398...$$

and was obtained in 2018 by Jenkinson and Pollicot. The lower bound we were able to obtain is a bit better than 0.344. Here is the table of the lower bounds coming from Equation (7) (values truncated, not rounded).

n	2	3	4	5	6	7	8	9	10
γ_n	0.34483	0.4121	0.44025	0.4553	0.4647	0.4711	0.4756	0.4790	0.4816

Let
$$E(\infty) = \bigcup_{m=1}^{\infty} E(\{1, 2, \dots, m\}).$$
The set $E(\infty)$ is very close to the set S we considered in Theorem 7.2.1. To be precise, $E(\infty) = \coprod_{a_0 \in \mathbb{N}^+} a_0 + S$. It is thus clear that they have the same dimension and the same measure
$$\mu(E(\infty)) = 0.$$

On the other hand, letting $m \to \infty$ in the lower estimate in (7), we find

$$\dim_H(E(\infty)) = \dim_H(S) \geq \frac{1}{2}.$$

This is rather satisfying, because we always believed S is large and complicated ☺

The truth is even more satisfying.

Theorem 8.6.1 (Jarník). [17] $\dim(S) = 1$.

We mention a very general and powerful theorem.

Theorem 8.6.2 (Ramharter). [18] *Let $A \subseteq \mathbb{N}^+$ be a subset, finite or infinite. Let*

$$E(A) = \{[a_0, a_1, a_2, \dots] : a_i \in A, \forall i\}.$$

Let μ and ν be the real numbers defined by

$$\sum_{m \in A} [\overline{m}]^{-\mu} = 1, \quad \sum_{m \in A} m^{-\nu} = 1,$$

where $[\overline{m}]$ denotes the periodic continued fraction $[m, m, m, \dots] = \frac{m+\sqrt{m^2+4}}{2}$. Then,

$$\mu \leq 2 \dim_H(R) \leq \nu.$$

Exercise 8.6.3. ★ Use Ramharter's Theorem to prove that $\dim_H(S) \geq 0.9$, say, where S is the set of all real numbers for which the partial quotients of their continued fractions take only finitely many values. The theorem is not powerful enough to imply that $\dim_H(S) = 1$, but it can be used for a great many examples of sets defined by conditions on continued fractions.

We have seen that the Cantor set $\mathscr{C}$ and the sets S_m in Theorem 7.2.1 have real presence; in spite of being of measure 0 they have positive Hausdorff dimension. Another such evidence are the following theorems:

Theorem 8.6.4. *Let $\mathscr{C}$ be the Cantor set then*

$$\mathscr{C} + \mathscr{C} = \{x + y : x, y \in \mathscr{C}\} = [0, 2].$$

Exercise 8.6.5. ★ Prove Theorem 8.6.4.

In contrast with the exercise, the next theorem is much harder.[19]

Theorem 8.6.6. *Let $M \geq 4$ be an integer and let $S_M = \{[0, a_1, a_2, \dots] : a_i \leq M, \forall i\}$. Then*

$$S_M + S_M = \left[\frac{-M + \sqrt{M^2 + 4M}}{M}, -M + \sqrt{M^2 + 4M}\right].$$

In particular, as the length of the interval $S_4 + S_4$ is greater than 1, this theorem implies the following:

Theorem 8.6.7 (M. Hall). *Any real number r can be written in the form*

$$r = n + [0, a_1, a_2, \dots] + [0, b_1, b_2, \dots], \quad n \in \mathbb{Z}, 1 \leq a_i, b_i \leq 4 \ \forall i.$$

[17] V. Jarník, Zur metrischen Theorie der diophantischen Approximationen, *Prace Matematyczno-Fizyczne* 36 (1928-1929), 91-106.

[18] See G. Ramharter, Some metrical properties of continued fractions. *Mathematika*, 30 (1983), 117–132. The statement here follows from Theorem 1, equation (40) and comments following it. For a more accessible proof, see T. W. Cusick, Continuants with bounded digits. III. *Monatsh. Math.* 99 (1985), no. 2, 105?109.

[19] The proof can be found in the book Rockett & Szüsz, *Continued Fractions*.

8.6.1. *The issue with* $E(\{1,2,\dots,m\})$. Finally, we come back to a point we left vague in the calculation above; namely, our claim that $E(\{1,2,\dots,m\})$ is the attractor for the IFS system considered above. The issue is that although $E(\{1,2,\dots,m\}) = \cup_{i=1}^{m} s_i(E(\{1,2,\dots,m\}))$, it is not immediately clear that it is a closed set. If it is, then it is the unique compact attractor. There are three options to resolve this issue.

The first is to show that indeed $E(\{1,2,\dots,m\})$ is closed, by showing that its complement is open. This is *not* immediately clear, as Falconer claims. A point θ in the complement is, for example, an irrational number one of whose partial quotients, say a_t, is bigger than m. It is then indeed not hard to see that any real number close enough to θ will also have the same a_t. To see that one uses that the sets we called $E\left(\begin{smallmatrix}1 & 2 & \cdots & t\\ k_1 & k_2 & \cdots & k_t\end{smallmatrix}\right)$ are intervals with rational end points and so θ cannot not be one of these end points.

However, a point θ in the complement could also be a rational point, which can very well be an end point of the aforementioned intervals, and the situation is then less clear. In this case one has to show that any r close enough to θ would have to have arbitrary large partial quotients and thus lie in the complement of $E(\{1,2,\dots,m\})$. That *is* true but requires an argument.

But, there is another option altogether. Let $A = E(\{1,2,\dots,m\})$ for ease of notation. First one argues that in our situation if $A = \cup_i s_i(A)$ then also $\bar{A} = \cup_i s_i(\bar{A})$, just because the s_i are bi-Lipschitz. And clearly $\bar{A} = \cup_i s_i(\bar{A})$ is the unique compact attractor.

Now if we show $A \subseteq \bar{A} \subseteq A \cup \mathbb{Q}$ then, since $\dim_H(A) = \dim_H(A \cup \mathbb{Q})$ (that follows easily from Theorem 8.1.4), also $\dim_H(A) = \dim_H(\bar{A})$ and we can just estimate $\dim_H(\bar{A})$. Now,

$$A = \cap_{n=1}^{\infty}\{[a_0, a_1, \dots, a_n, \dots] : 1 \le a_i \le m, i = 1,2,\dots,n\} \tag{8}$$

and each set $\{[a_0, a_1, \dots, a_n, \dots] : 1 \le a_i \le m, i = 1,2,\dots,n\}$ is a finite union of sets of the form $E\left(\begin{smallmatrix}1 & 2 & \cdots & t\\ k_1 & k_2 & \cdots & k_t\end{smallmatrix}\right)$ and so $\{[a_0, a_1, \dots, a_n, \dots] : 1 \le a_i \le m, i = 1,2,\dots,n\}$ is contained in a finite union of closed intervals with rational end points. Thus, we can replace the intersection (8) by an intersection of sets that are each a finite union of closed intervals with rational end points. That larger intersection is closed. What we have added, in the end, is at most some rational points (that arise as end points of intervals) to our set A.

The third option is to check that $E = \cap_{k=0}^{\infty} S^k(D)$, directly from the definition of E and the description of the s_i in terms of continued fractions. Since each $S^k(D)$ is closed, it follows that E is closed too.

9. In conclusion

The subject of continued fractions is rich and has many applications to different areas of mathematics. Developing functions into continued fractions is a subject that we did not really discuss, but it is a subject that is both ancient and of current research in optimization and numerical analysis. There are many other generalizations of continued fractions, but even the simplest kind of continued fractions, those that we have discussed in detail, offer many interesting problems and continue to be a very active field of research. In particular, their applications to transcendence theory and to dynamical systems are still actively researched. The MathSciNet search for "continued fractions in title" returns 3650 hits on January 2020, out of which 391 are from 2016 or later.

In spite of all this progress there are some outstanding problems just waiting for a brilliant solution. For example, recall the set S we studied in Theorem 7.2.1, the set of all real numbers x in $(0,1)$ whose partial quotients are bounded by some bound, which may depend on x. It is conjectured that any number in S that is not rational, or quadratic, is transcendental. If you are interested in learning more about this aspect, a good place to start may be the article B. Adamczewski, Y. Bugeaud and L. Davison, "Continued fractions and transcendental numbers. Numération, pavages, substitutions." *Ann. Inst. Fourier (Grenoble)* 56 (2006), no. 7, 2093–2113.

EQUATIONS OVER FINITE FIELDS

10. INTRODUCTION

In this chapter we follow closely

Kenneth Ireland and Michael Rosen, *A classical introduction to modern number theory.* Second edition.

It is an excellent book, but the material required to prove the main results we are interested in is spread over many of its chapters. The exposition we offer here is a more concise path towards those goals than the book itself. For an introductory text about algebraic geometry, we recommend

William Fulton, *Algebraic curves. An introduction to algebraic geometry.*

Our main topic in this chapter is the study of the number of solutions to a system of polynomial equations over a finite field. Let $\mathbb{F}$ be a finite field, for example $\mathbb{F}_p$, and suppose we are given m polynomials with coefficients in $\mathbb{F}$ in n variables, $f_1(x_1,\dots,x_n),\dots,f_m(x_1,\dots,x_n)$. We are interested in counting the solutions to the system of equations:

$$V: \quad \begin{cases} f_1(x_1,\dots,x_n)=0, \\ \vdots \\ f_m(x_1,\dots,x_n)=0. \end{cases}$$

Namely, we want to find

$$\sharp V(\mathbb{F}) = \sharp\{(a_1,\dots,a_n)\in\mathbb{F}^n : f_i(a_1,\dots,a_n)=0, \forall i=1,\dots,m\}.$$

If $\mathbb{L}$ is a field containing $\mathbb{F}$, then it likewise makes sense to ask about

$$\sharp V(\mathbb{L}) = \sharp\{(a_1,\dots,a_n)\in\mathbb{L}^n : f_i(a_1,\dots,a_n)=0, \forall i=1,\dots,m\}.$$

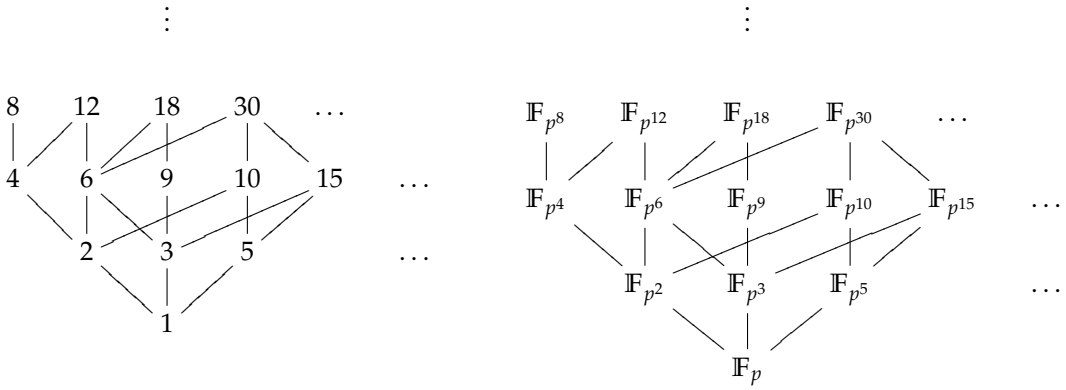
Let us make our discussion more systematic by carefully explaining which fields $\mathbb{L}$ we want to consider. Recall that for two fields $\mathbb{F}\subseteq\mathbb{L}$ one defines $[\mathbb{L}:\mathbb{F}]=\dim_{\mathbb{F}}(\mathbb{L})$, which is either a positive integer or ∞. It is called the **degree** of $\mathbb{L}$ over $\mathbb{F}$.

Let p be a prime and let $\mathbb{F}_p$ be the field of p elements; $\mathbb{F}_p\cong\mathbb{Z}/p\mathbb{Z}$. Let $\overline{\mathbb{F}}_p$ be an algebraic closure of $\mathbb{F}_p$ (see §11.1 for a more thorough discussion). Then: (1) for every integer n , there is a unique subfield of $\overline{\mathbb{F}}_p$ with p^n elements that we denote $\mathbb{F}_{p^n}$. (2) We have an inclusion $\mathbb{F}_{p^m}\subseteq\mathbb{F}_{p^n}$ if and only if $m|n$, and then $[\mathbb{F}_{p^n}:\mathbb{F}_{p^m}]=n/m$. (3) $\overline{\mathbb{F}}_p=\cup_{n=1}^{\infty}\mathbb{F}_{p^n}$.

It is a fact that every finite field of p^n elements is isomorphic to $\mathbb{F}_{p^n}$. Thus, to study finite fields of characteristic p, we might as well restrict our attention to subfields of $\overline{\mathbb{F}}_p$.

It follows from the above that if $\mathbb{F}$ is any finite field contained in $\overline{\mathbb{F}}_p$, then for every positive integer s there is a unique field $\mathbb{F}_{[s]}\subset\overline{\mathbb{F}}_p$ containing $\mathbb{F}$, such that $[\mathbb{F}_{[s]}:\mathbb{F}]=s$. Otherwise said, the lattice of finite subfields of $\overline{\mathbb{F}}_p$ is exactly like the lattice of positive integers $\mathbb{N}^+$ with the divisibility relation.

(9)



As above, let $f_i(x_1, \ldots, x_n), i = 1, \ldots, m$, be polynomials in n variables with coefficients in a finite field $\mathbb{F} \subset \overline{\mathbb{F}}_p$. Let $V(\mathbb{F}_{[s]})$ denote the set of solutions in $\mathbb{F}_{[s]}$:

$$V(\mathbb{F}_{[s]}) = \{(a_1, \ldots, a_n) :\in \mathbb{F}_{[s]}^n : f_j(a_1, \ldots, a_n) = 0, j = 1, \ldots, m.\}$$

One is interested in $\sharp V(\mathbb{F}_{[s]})$. This quantity is obviously of interest to number theory and algebraic geometry, but also to cryptography, coding theory and combinatorics, to name a few other areas. Pierre Deligne got his Fields Medal in 1978 for proving very precise information about the behaviour of the numbers $\sharp V(\mathbb{F}_{[s]})$.

It turns out that it is better to work in a projective space and with homogeneous polynomial equations defining smooth, i.e. non-singular, projective varieties V over $\mathbb{F}$. We discuss all these concepts below, but let us assume for time being that we know what they mean. We define the **zeta function** of V, $\zeta_V(T)$, as the series

$$\zeta_V(T) = \exp\left(\sum_{s=1}^{\infty} \frac{\sharp V(\mathbb{F}_{[s]})}{s} \cdot T^s\right). \tag{10}$$

This looks a bit intimidating at first sight, but note that the zeta function is designed so that

$$\mathrm{d}\log\ \zeta_V(T) = \sum_{s=1}^{\infty} \sharp V(\mathbb{F}_{[s]}) \cdot T^{s-1}.$$

So, $\zeta_V(T)$ is a jazzed-up version of a generating series for the sequence of integers

$$\sharp V(\mathbb{F}_{[s]}),\ \ s = 1, 2, 3, \ldots$$

As usual in math, the definitions are justified by the theorems we can prove about them.

10.1. **Weil's conjectures.** In 1949, André Weil made his conjectures about $\zeta_V(T)$ in a short but seminal article[20]; an article that influenced greatly the development of algebraic geometry and number theory. Some parts of his conjectures were proven by A. Grothendieck ("Functional equation") and B. Dwork ("Rationality"), but the hardest part ("Riemann hypothesis") was proven by Deligne.

[20] A. Weil, "Numbers of solutions of equations in finite fields", *Bulletin of the American Mathematical Society*, 55 (5): 497–508.

Weil's conjectures. Assume that V is a non-singular irreducible projective variety of dimension d over a finite field $\mathbb{F}_q$ with q elements.

(1) (*Rationality*) The function $\zeta_V(T)$ is a ratio of polynomials with rational coefficients.

$$\zeta_V(T) = \frac{P_1(T)P_3(T)\cdots P_{2d-1}(T)}{P_0(T)P_2(T)\cdots P_{2d}(T)} \in \mathbb{Q}(T).$$

In fact, for every i, $P_i(T) \in \mathbb{Z}[T]$ and $P_i(0) = 1$. We always have,

$$P_0(T) = 1 - T, \quad P_{2d}(T) = 1 - q^d T.$$

(2) (*Functional equation*) For a suitable integer E[21], and a suitable sign $\epsilon \in \{\pm 1\}$, we have

$$\zeta_V(q^{-d}T^{-1}) = \epsilon q^{dE/2} T^E \zeta_V(T).$$

(3) (*Riemann hypothesis*) Write

$$P_i(T) = \prod_j (1 - \alpha_{ij} T), \quad \alpha_{ij} \in \mathbb{C}.$$

For all i, j

$$|\alpha_{ij}| = q^{i/2}.$$

(4) (*Cohomological interpretation*) This part provides information about the degrees of P_i in terms of the cohomology of V. We will make an effort in §16 to give some idea, even if rather vague, regarding what this is all about.

An important consequence of the Weil conjectures is that they provide precise information about the numbers $\sharp V(\mathbb{F}_{[s]})$. Indeed,

$$\begin{aligned}
\sum_{s=1}^{\infty} \sharp V(\mathbb{F}_{[s]}) \cdot T^{s-1} &= \mathrm{d}\log\ \zeta_V(T) \\
&= \sum_{i=0}^{2d} (-1)^{i+1}\ \mathrm{d}\log\ P_i(T) \\
&= \sum_{i=0}^{2d} (-1)^{i+1} \sum_j\ \mathrm{d}\log\ (1 - \alpha_{ij}(T)) \\
&= \sum_{i=0}^{2d} (-1)^{i} \sum_j \frac{\alpha_{ij}}{1 - \alpha_{ij} T} \\
&= \sum_{i=0}^{2d} (-1)^{i} \sum_j \alpha_{ij}(1 + \alpha_{ij} T + \alpha_{ij}^2 T^2 + \dots).
\end{aligned} \tag{11}$$

Thus, the knowledge of the α_{ij} gives us formulas for $\sharp V(\mathbb{F}_{[s]})$ for every s. In particular,

$$\sharp V(\mathbb{F}) = \sum_{i,j} (-1)^i \alpha_{ij}.$$

[21] E is the Euler characteristic of V, a certain cohomological invariant of V that we do not describe here.

11. Some prerequisites

We provide some background here concerning finite fields, projective space, affine and projective varieties.

11.1. **Finite fields - a short summary.** Let p be a prime and let $\mathbb{F}_p$ be a field with p elements. For example, $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$. In fact, any such $\mathbb{F}_p$ is (uniquely) isomorphic to $\mathbb{Z}/p\mathbb{Z}$.

(1) Every finite field $\mathbb{F}$ contains some $\mathbb{F}_p$; p is unique and is called the **characteristic** of $\mathbb{F}$. In $\mathbb{F}$, we have
$$p := \underbrace{1 + \cdots + 1}_{p-\text{times}} = 0,$$
and consequently, for every $z \in \mathbb{F}$, $pz := z + \cdots + z = (1 + \cdots + 1)z = 0$ (the sums are p-times). Since for $1 \leq i \leq p-1$ one has $p|\binom{p}{i}$, we deduce from the binomial formula that we have $(x+y)^p = x^p + y^p$, and, by iterating, for any power $q = p^s$ of p we have
$$(x+y)^q = x^q + y^q, \quad q = p^s.$$
(2) Let $[\mathbb{F} : \mathbb{F}_p] = s$ then $\sharp\mathbb{F} = p^s =: q$. If $\mathbb{L} \supseteq \mathbb{F}$ is a finite field and $t = [\mathbb{L} : \mathbb{F}]$ then $\sharp\mathbb{L} = q^t$.
(3) The set $\mathbb{F}^\times = \mathbb{F} \setminus \{0\}$ is a cyclic group of order $q-1$ under multiplication. Every element in $\mathbb{F}$ is a solution of $x^q - x = 0$. Thus, if $\overline{\mathbb{F}}_p$ is the algebraic closure of $\mathbb{F}_p$ then, for every s, there is a unique subfield of $\overline{\mathbb{F}}_p$ with $q = p^s$ elements. Namely, the set of solutions of the polynomial $x^q - x = 0$ in $\overline{\mathbb{F}}_p$. Moreover, any two fields with q elements are isomorphic.
(4) The lattice of subfields of $\overline{\mathbb{F}}_p$ is the same as the lattice of positive integers with division. $\mathbb{F}_{p^s} \leftrightarrow s$, $\mathbb{F}_{p^s} \subseteq \mathbb{F}_{p^{s_1}} \Leftrightarrow s|s_1$; cf. the diagram in (9).
(5) The **Frobenius** map
$$\varphi \colon \overline{\mathbb{F}}_p \to \overline{\mathbb{F}}_p, \quad \varphi(x) = x^p,$$
is an automorphism of fields. That is, it is bijective and $\varphi(x+y) = \varphi(x) + \varphi(y)$ (by the binomial formula in $\mathbb{F}$), $\varphi(xy) = \varphi(x)\varphi(y)$, and $\varphi(1) = 1$. We have
$$\mathbb{F}_{p^s} = \{x \in \overline{\mathbb{F}}_p : \varphi^s(x) := \underbrace{(\varphi \circ \cdots \circ \varphi)}_{s-\text{times}}(x) = x^{p^s} = x.\}$$
(6) Let $\mathbb{F} = \mathbb{F}_{p^s} = \mathbb{F}_q$ and $\mathbb{L} = \mathbb{F}_{[t]} = \mathbb{F}_{q^t}$. Define the **trace map**
$$\mathrm{Tr}_{\mathbb{L}/\mathbb{F}}(x) = x + x^q + \ldots x^{q^{t-1}},$$
and the **norm map**
$$\mathrm{Nm}_{\mathbb{L}/\mathbb{F}}(x) = x \cdot x^q \cdots x^{q^{t-1}}.$$
Then,
- $\mathrm{Tr}_{\mathbb{L}/\mathbb{F}}(x)$ is an $\mathbb{F}$-linear surjective map
$$\mathrm{Tr}_{\mathbb{L}/\mathbb{F}} \colon \mathbb{L} \to \mathbb{F}.$$
- $\mathrm{Nm}_{\mathbb{L}/\mathbb{F}}(x)$ is a surjective homomorphism of groups
$$\mathrm{Nm}_{\mathbb{L}/\mathbb{F}} \colon \mathbb{L}^\times \to \mathbb{F}^\times.$$
Indeed, it is not hard to prove that the trace is additive and the norm is multiplicative. To prove that the image lies in $\mathbb{F}$ in both cases, it is enough to prove that
$$\varphi^s(x + x^q + \ldots x^{q^{t-1}}) = x + x^q + \ldots x^{q^{t-1}}, \quad \varphi^s(x \cdot x^q \cdots x^{q^{t-1}}) = x \cdot x^q \cdots x^{q^{t-1}}.$$
Note that $\varphi^s(x) = x^q$, and $x^{q^t} = x$, for $x \in \mathbb{L}$, so the statements about the images are true. Since, for $f \in \mathbb{F}$ we have $f^q = f$, it follows that the trace is linear over $\mathbb{F}$. The only remaining point is the surjectivity or the trace and norm, which we leave as an exercise.

Exercise 11.1.1. Prove that the trace $\mathrm{Tr}_{\mathbb{L}/\mathbb{F}}$ and norm $\mathrm{Nm}_{\mathbb{L}/\mathbb{F}}$ are surjective maps onto $\mathbb{F}$ and $\mathbb{F}^\times$, respectively.

11.2. **Projective space.** Let $\mathbb{F}$ be any field. We denote elements of $\mathbb{F}^{n+1}$ by vectors $(x_0, \dots, x_n)$. We define an equivalence relation on $\mathbb{F}^{n+1} \setminus \{0\}$ by decreeing that for any $\alpha \in \mathbb{F}^\times$,

$$(x_0, \dots, x_n) \sim (\alpha x_0, \dots, \alpha x_n).$$

We denote by

$$[x_0 : \cdots : x_n]$$

the equivalence class of $(x_0, \dots, x_n)$.

As $(x_0, \dots, x_n)$ is not the zero vector, it defines a line through the origin in $\mathbb{F}^{n+1}$; to wit,

$$\{\alpha \cdot (x_0, \dots, x_n) : \alpha \in \mathbb{F}\}.$$

Any $(y_0, \dots, y_n)$ equivalent to $(x_0, \dots, x_n)$ defines the same line, and vice-versa. Thus, the equivalence classes are in bijection with lines through the origin in $\mathbb{F}^{n+1}$. For example, for $\mathbb{P}^1(\mathbb{R})$ we have the following diagram, showing that $\mathbb{P}^1(\mathbb{R})$ is the quotient of the circle by the automorphism $x \mapsto -x$.

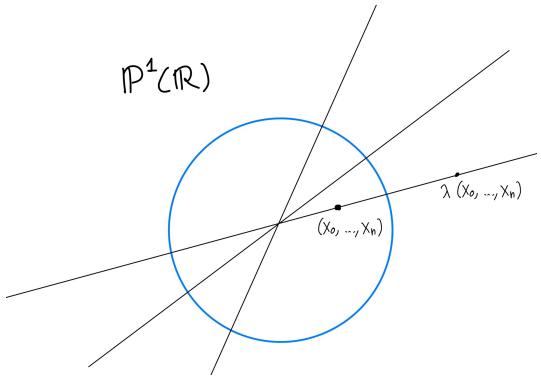


FIGURE 10. The projective line over $\mathbb{R}$.

The collection of equivalence classes is called the n-dimensional **projective space** over $\mathbb{F}$:

$$\mathbb{P}^n(\mathbb{F}) = \{[x_0 : \cdots : x_n] : x_i \in \mathbb{F} \text{ not all zero}\}.$$

Lemma 11.2.1. *There is a natural partition*

$$\mathbb{P}^n(\mathbb{F}) = \mathbb{F}^n \coprod \mathbb{F}^{n-1} \coprod \cdots \coprod \mathbb{F}^0.$$

Proof. We prove that by induction on n. For $n = 0$, $\mathbb{P}^0(\mathbb{F})$ consists of just one point so $\mathbb{P}^0(\mathbb{F}) = \mathbb{F}^0$, where $\mathbb{F}^0$ is the 0-dimensional vector space over $\mathbb{F}$, but we just think about it as a set with 1 point. For $n = 1$ we have

$$\begin{aligned}\mathbb{P}^1(\mathbb{F}) &= \{[x_0 : x_1] : x_1 \neq 0\} \coprod \{[x_0 : 0] : x_0 \neq 0\} \\ &= \{[y_0 : 1] : y_0 \in \mathbb{F}\} \coprod \{[1 : 0]\} \\ &= \mathbb{F} \coprod \mathbb{F}^0.\end{aligned}$$

The point of this calculation is that $[1:0] = [x_0:0]$ for all $x_0 \neq 0$ and $[y_0:1] = [\alpha y_0:\alpha]$ for all $\alpha \neq 0$. Now, inductively,

$$\begin{aligned}\mathbb{P}^n(\mathbb{F}) &= \{[x_0:x_1:\cdots:x_n] : x_n \neq 0\} \qquad \coprod \{[x_0:\cdots:x_{n-1}:0] \in \mathbb{P}^n(\mathbb{F})\} \\ &= \{[y_0:y_1:\cdots:y_{n-1}:1] : y_i \in \mathbb{F}\} \coprod \{[x_0:\cdots:x_{n-1}] \in \mathbb{P}^{n-1}(\mathbb{F})\} \\ &= \mathbb{F}^n \coprod \mathbb{P}^{n-1}(\mathbb{F}) \\ &= \mathbb{F}^n \coprod \mathbb{F}^{n-1} \coprod \cdots \coprod \mathbb{F}^0.\end{aligned}$$

□

On the **affine space** $\mathbb{F}^n$, which we often denote $\mathbb{A}^n(\mathbb{F})$, we can define functions using polynomial functions $f(x_1,\ldots,x_n) \in \mathbb{F}[x_1,\ldots,x_n]$. On the other hand, if we consider any polynomial in $f(x_0,x_1,\ldots,x_n) \in \mathbb{F}[x_0,x_1,\ldots,x_n]$ and try to assign it a value at an element $[a_0:\cdots:a_n] \in \mathbb{P}^n(\mathbb{F})$ by saying that the value is $f(a_0,a_1,\ldots,a_n)$, this is not well defined - the value depends on the representative $(a_0,a_1,\ldots,a_n)$ for $[a_0:a_1:\cdots:a_n]$. For example, take $f(x_0,x_1) = x_0+1$. Then, $f(1,1) = 2$. But the point $[1:1] = [2:2]$ in $\mathbb{P}^1(\mathbb{F})$ and $f(2,2) = 3$. So, evaluating polynomials on $\mathbb{P}^n(\mathbb{F})$ is not a well defined operation.

Let us then restrict our attention to **homogeneous polynomials**. Namely, polynomials that are the sum of monomials of the same degree, For example $x_0x_1^2 + 3x_0x_1x_2$ is homogeneous polynomial of degree 3. Let d be a non-negative integer and consider $I = (i_0,\ldots,i_n)$, a vector of non-negative integers such that $|I| := i_0+i_1+\cdots+i_n = d$. Denote $x^I = x_0^{i_0}x_1^{i_1}\cdots x_n^{i_n}$, which is a monomial of degree d. Then, the general form of a homogeneous polynomial over $\mathbb{F}$ of degree d is

$$f(x_0,x_1,\ldots,x_n) = \sum_{I=(i_0,\ldots,i_n),|I|=d} a_I x^I,$$

where $a_I \in \mathbb{F}$ and the summation is over all I with non-negative integer coordinates, such that $|I| = d$. Now, although the value of such an f at a point $[a_0:\cdots:a_n] \in \mathbb{P}^n(\mathbb{F})$ is still not well-defined, the statement $f(a_0,\ldots,a_n) = 0$ *is* well-defined. That is, if f is homogeneous of degree d then $f(\alpha a_0,\ldots,\alpha a_n) = \alpha^d f(a_0,\ldots,a_n)$. And, as $\alpha \neq 0$, whether $f(a_0,\ldots,a_n) = 0$, or not, is a statement independent of the particular representative $(a_0,\ldots,a_n)$ for the equivalence class $[a_0:\cdots:a_n]$ chosen for the evaluation of f.

11.3. **Affine and projective varieties.** Polynomials and homogeneous polynomials can be used to define algebraic varieties in affine space and projective space. We define these notions and discuss the connection between them.

11.3.1. *Affine varieties.* Let $\mathbb{F}$ be a field. We change slightly our notation for $\mathbb{A}^n(\mathbb{F})$, writing

$$\mathbb{A}^n(\mathbb{F}) = \mathbb{F}^n = \{(a_0,\ldots,a_{n-1}) : a_i \in \mathbb{F}\}.$$

Let $f_i(x_0,\ldots,x_{n-1})$, $i = 1,\ldots,m$, be polynomials in the n-variables $x_0,\ldots,x_{n-1}$. The **affine variety** $V = Z(f_1,\ldots,f_m)$ is defined by the vanishing of all $f_1,\ldots,f_m$. Namely, it has points

$$V(\mathbb{F}) = \{(a_0,\ldots,a_{n-1}) \in \mathbb{F}^n : f_i(a_0,\ldots,a_{n-1}) = 0, i = 1,\ldots,m\}.$$

Note that if $\mathbb{L}$ is any field containing $\mathbb{F}$, it makes sense to talk about

$$V(\mathbb{L}) = \{(a_0,\ldots,a_{n-1}) \in \mathbb{L}^n : f_i(a_0,\ldots,a_{n-1}) = 0, i = 1,\ldots,m\}.$$

So, in a sense that we will not formalize here, the variety V is more than just its points $V(\mathbb{F})$.

Such a variety V has a well-defined notion of **dimension**. Unfortunately, it is a bit tricky to define, so we will not do that here. It is defined in Fulton's book, and in any comprehensive

text about algebraic geometry. Often the dimension is just $n-m$, where m is the number of polynomial equations used to define V in $\mathbb{A}^n$; this happens if the equations are "sufficiently generic", but this is not always the case.

The variety V is called **non-singular**, or **smooth**, if for every $a=(a_0,\dots,a_{n-1})\in V(\mathbb{L})$, for any $\mathbb{L}\supset\mathbb{F}$, the rank of the matrix

$$\left(\frac{\partial f_i}{\partial x_j}(a)\right)_{i,j}$$

is $n-d$, where d is the dimension of V.

In fact, mostly we will interested in varieties V defined by a single non-constant polynomial equation $f(x_0,\dots,x_{n-1})$ of degree d and V is just the solutions to this polynomial. Such a variety is called a **hypersurface** of degree d. Many of the concepts mentioned above simplify in this case: the dimension of V is always $n-1$ and it is non-singular if and only if the vector

$$\left(\frac{\partial f}{\partial x_0}(a),\dots,\frac{\partial f}{\partial x_{n-1}}(a)\right)\neq 0$$

for any $a\in V$. Hopefully, you have seen the exact same criterion in the context of manifolds in a course in calculus or differential geometry.

For every field $\mathbb{L}$, $V(\mathbb{L})$ is the solutions to this polynomial in $\mathbb{L}$. Consider for example the equation $x_1^2=x_0^3+ax_0+b$, where $a,b\in\mathbb{F}$ are some constants. The variety V it defines, if non-singular, is called an **elliptic curve**.

11.3.2. *Projective varieties.* Let $\mathbb{F}$ be a field. Let $f_i(x_0,\dots,x_n)$, $i=1,2,\dots,m$, be homogeneous polynomials, possibly of different degrees. They define a **projective variety** V such that for every field $\mathbb{L}\supseteq\mathbb{F}$,

$$V(\mathbb{L})=\{a=[a_0:a_1:\cdots:a_n]\in\mathbb{P}^n(\mathbb{L}):f_i(a)=0,\ i=1,2,\dots,m\}.$$

We are using the same notation V as in the affine case. If we need to distinguish between the two we will use V^{aff} and V^{proj}. Note that

$$V^{\mathrm{proj}}=(V^{\mathrm{aff}}-\{0\})/\sim.$$

That is, the projective variety defined by $f_1,\dots,f_m$ in $\mathbb{P}^n(\mathbb{F})$ is the equivalence classes of the non-zero points of the affine variety defined by $f_1,\dots,f_m$ in $\mathbb{A}^{n+1}(\mathbb{F})$.

Let $f(x_0,\dots,x_{n-1})\in\mathbb{F}[x_0,\dots,x_{n-1}]$. We define the **homogenization** of f, denoted $f^{[h]}$, to be the homogeneous polynomial in $\mathbb{F}[x_0,\dots,x_{n-1},x_n]$ given by the formula:

$$f^{[h]}(x_0,x_1,\dots,x_n)=x_n^{\deg(f)}f\left(\frac{x_0}{x_n},\dots,\frac{x_{n-1}}{x_n}\right).$$

In practice, this is a simple operation: just add a suitable power of x_n to each monomial of f so as to get a homogeneous polynomial of degree equal $\deg(f)$. For example, if $f(x_0,x_1)=x_0^2-x_1^3-ax_1-b$ then $f^{[h]}(x_0,x_1,x_2)=x_0^2x_2-x_1^3-ax_1x_2^2-bx_2^3$. Note that

$$f^{[h]}(x_0,\dots,x_{n-1},1)=f(x_0,\dots,x_{n-1}).$$

From this we get the following:

$$\begin{aligned}V^{\mathrm{aff}}(f_1,\dots,f_m)&=\{(a_0,\dots,a_{n-1})\in\mathbb{A}^n:f_j(a_0,\dots,a_{n-1})=0,\forall j\}\\ V^{\mathrm{proj}}(f_1^{[h]},\dots,f_m^{[h]})&=\{[a_0:\cdots:a_{n-1}:a_n]\in\mathbb{P}^n:f_j^{[h]}(a_0,\dots,a_{n-1},a_n)=0,\forall j\}.\end{aligned}$$

And, under the decomposition $\mathbb{P}^n = \mathbb{A}^n \coprod \mathbb{P}^{n-1}$ of Lemma 11.2.1, we have

$$V^{\text{proj}}(f_1^{[h]}, \ldots, f_m^{[h]}) \cap \mathbb{A}^n = V^{\text{aff}}(f_1, \ldots, f_m).$$

Therefore, $V^{\text{proj}}(f_1^{[h]}, \ldots, f_m^{[h]})$ is a projective variety extending $V^{\text{aff}}(f_1, \ldots, f_m)$. [22]

For example, consider the elliptic curve $f(x_0, x_1) = x_0^2 - x_1^3 - ax_1 - b$ and $f^{[h]}(x_0, x_1, x_2) = x_0^2 x_2 - x_1^3 - ax_1 x_2^2 - bx_2^3$.

- If $x_2 \neq 0$, we may assume $x_2 = 1$ and the solutions for $f^{[h]}$ are $[x_0 : x_1 : 1]$ where (x_0, x_1) are solutions to f.
- If $x_2 = 0$, we get one additional solution, one not lying in $\mathbb{A}^2$, which is the point $[1 : 0 : 0]$.

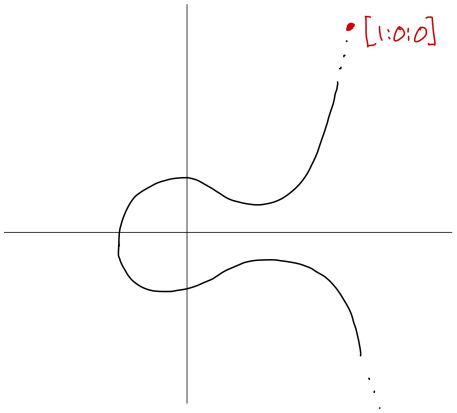


FIGURE 11. An elliptic curve in $\mathbb{P}^2$.

Finally, a projective variety V of dimension d in $\mathbb{P}^n$ (coming from the affine variety of dimension $d+1$ in $\mathbb{A}^{n+1}$ defined by the same polynomials) is **non-singular**, or **smooth**, if the rank of the following matrix is $n - d$ at every point $a = (a_0, \ldots, a_n) \neq (0, 0, \ldots, 0)$ lying on V:

$$\left(\frac{\partial f_i}{\partial x_j}(a) \right)_{i,j}.$$

Thus, $V^{\text{proj}}(f_1, \ldots, f_m) \subseteq \mathbb{P}^n$ is smooth if and only if $V^{\text{aff}}(f_1, \ldots, f_m) \subseteq \mathbb{A}^{n+1}$ is smooth, except possibly at $(0, 0, \ldots, 0)$.

In fact, in most of this chapter we will be concerned with the following **key example**. Let $\mathbb{F}$ be a field of characteristic p. Suppose that $p \nmid m$ and $\alpha_0, \ldots, \alpha_n \in \mathbb{F}$ are all non-zero. Let

$$f(x_0, \ldots, x_n) = \alpha_0 x_0^m + \cdots + \alpha_n x_n^m.$$

The projective variety V that it defines is a hypersurface of degree m and dimension $n-1$. It is non-singular, because

$$\left(\frac{\partial f}{\partial x_0}, \frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right) = (m\alpha_0 x_0^{m-1}, m\alpha_1 x_1^{m-1}, \ldots, m\alpha_n x_n^{m-1})$$

is zero at a point $(a_0, \ldots, a_m)$ implies that $a_i = 0$ for all i. But, the point $[0 : 0 : \cdots : 0]$ is not a point in the projective space.

Our goal is to understand the zeta function of V.

[22] A word of caution: $V^{\text{proj}}(f_1^{[h]}, \ldots, f_m^{[h]})$ need not be the *minimal* projective variety containing $V^{\text{aff}}(f_1, \ldots, f_m)$.

12. Some examples of zeta functions

Before continuing to develop tools to study the number of points on hypersurfaces, let us look at some very particular examples. One of those examples, the Grassmannian, was in fact among the examples Weil originally provided for his general conjectures.

12.1. **The zeta function of $\mathbb{P}^n$.** The projective space $\mathbb{P}^n$ considered over $\mathbb{F}_p$ is the simplest projective variety. It is defined by the empty set of polynomials, or, if you wish, by the zero polynomial, and it is smooth. The decomposition of Lemma 11.2.1 gives us

$$\sharp\mathbb{P}^n(\mathbb{F}_{p^s}) = 1 + p^s + \cdots + p^{ns} = \frac{1-p^{(n+1)s}}{1-p^s}.$$

Before calculating $\zeta_{\mathbb{P}^n}$ we note the following identities:

$$\exp\left(\sum_{s=1}^{\infty} x^s \frac{T^s}{s}\right) = \exp\left(\sum_{s=1}^{\infty} \frac{(xT)^s}{s}\right) = \exp\left(-\log(1-xT)\right) = \frac{1}{1-xT},$$

where we used $\log(1+t) = t - \frac{t^2}{2} + \frac{t^3}{3} - \frac{t^4}{4} + \ldots$.

Now,

$$\begin{aligned}\zeta_{\mathbb{P}^n}(T) &= \exp\left(\sum_{s=1}^{\infty}(1+p^s+\cdots+p^{sn})\frac{T^s}{s}\right)\\ &= \prod_{j=0}^{n}\exp\left(\sum_{s=1}^{\infty}(p^j)^s\frac{T^s}{s}\right)\\ &= \frac{1}{(1-T)(1-pT)\cdots(1-p^nT)}.\end{aligned}$$

12.2. **The zeta function of the Grassmann variety $G_{m,n}$.** Recall that the projective space $\mathbb{P}^{n-1}$ has the interpretation as parametrizing lines through the origin in $\mathbb{A}^n$. Let $0 < m < n$ be an integer. For a given field $\mathbb{L}$ we may want to parameterize all the m-dimensional subspaces of $\mathbb{A}^n$. The case $m = 1$ is solved by the projective space - an algebraic variety such that for every $\mathbb{L}$, $\mathbb{P}^{n-1}(\mathbb{L})$ is in natural bijection with the 1-dimensional subspaces (lines through the origin) in $\mathbb{A}^n(\mathbb{L})$. It turns out, and this is one of the most basic constructions in algebraic geometry, that for every m there is a projective variety $G_{m,n}$, called a **Grassmannian**, that parameterizes m-dimensional subspaces of $\mathbb{A}^n$; that is, for every field $\mathbb{L}$, $G_{m,n}(\mathbb{L})$ is in natural bijection with m-dimensional subspaces of $\mathbb{A}^n(\mathbb{L}) = \mathbb{L}^n$. Moreover, $G_{m,n}$ is non-singular of dimension $m(n-m)$. The case $G_{1,n}$ is just $\mathbb{P}^{n-1}$.

Let $P = P_{m,n}$ be the subgroup of GL_n such that $P(\mathbb{L})$ consists of matrices M of the form

$$M = \begin{pmatrix} A & B \\ 0 & D \end{pmatrix}, \quad A \in \mathrm{GL}_m(\mathbb{L}), B \in M_{m,n-m}(\mathbb{L}), D \in \mathrm{GL}_{n-m}(\mathbb{L}).$$

Exercise 12.2.1. Show that for a field $\mathbb{L}$ there is a natural bijection

$$\mathrm{GL}_n(\mathbb{L})/P_{m,n}(\mathbb{L}) \leftrightarrow G_{m,n}(\mathbb{L}).$$

This bijection is associating to a right coset $M \cdot P_{m,n}(\mathbb{L})$ the m-dimensional subspace of $\mathbb{A}^n$ spanned by the first m-columns of M.

Using the bijection of the previous exercise, if $\mathbb{L}$ is a finite field we can count the number of points in $G_{m,n}(\mathbb{L})$.

Exercise 12.2.2. Let $\mathbb{L}$ be a finite field with q elements. Prove that

$$\sharp \mathrm{GL}_n(\mathbb{L}) = \prod_{i=1}^{n} (q^n - q^{n-i}) =: c(n),$$

$$\sharp G_{m,n}(\mathbb{L}) = \frac{c(n)}{c(m)c(n-m)q^{m(n-m)}}.$$

Verify the formula for the case of the projective space.

In general, writing the zeta function of $G_{m,n}$ based on this formula requires some more combinatorics, in particular the introduction of the Gaussian binomial coefficients. Let us then only state one more example before moving on.

Exercise 12.2.3. Prove that $G_{2,4}$, considered as a variety over $\mathbb{F}_p$, has the following zeta function:

$$\zeta(T) = \frac{1}{(1-T)(1-pT)(1-p^2T)^2(1-p^3T)(1-p^4T)}.$$

As another example, unrelated to Grassmannians, you may try and solve the following.

Exercise 12.2.4. Calculate the zeta function of the projective surface $x_0x_1 - x_2x_3 = 0$ over $\mathbb{F}_p$, in $\mathbb{P}^3$.

12.3. **Elliptic curves.** Let $p > 2$ be a prime and let q a power of p. Consider an elliptic curve, say $y^2 = x^3 + ax + b$, $a, b \in \mathbb{F} := \mathbb{F}_q$, or, rather, its projective model in $\mathbb{P}^2$ with coordinates x, y, z:

$$E: \quad y^2z - (x^3 + axz^2 + bz^3) = 0.$$

If we think about x as chosen randomly from $\mathbb{F}_q$, there are q such choices and we don't see any compelling reason for $x^3 + ax + b$ to be a square, or not be a square, in $\mathbb{F}_q$. Thus, let us accept that with probability about $1/2$ the quantity $x^3 + ax + b$ is a square in $\mathbb{F}_q$ and then, unless it is zero, there will be two $y \in \mathbb{F}_q$ such that $y^2 = x^3 + ax + b$. This heuristic suggests that

$$\sharp E(\mathbb{F}_q) \sim q.$$

Let us see what the Weil conjectures predict. We have

$$\zeta_E(T) = \frac{P_1(T)}{(1-T)(1-qT)}.$$

The cohomological interpretation tells us in this case that $\deg(P_1) = 2$. Write then

$$P_1(T) = (1-\alpha T)(1-\beta T), \quad \alpha, \beta \in \mathbb{C}, \quad |\alpha| = |\beta| = \sqrt{q}.$$

Referring to our calculations in Equation (11), we have

$$\sum_{s=1}^{\infty} \sharp E(\mathbb{F}_{[s]}) \cdot T^{s-1} = \sum_{n=0}^{\infty} t^n + \sum_{n=0}^{\infty} q^{n+1}t^n - \sum_{n=0}^{\infty} \alpha^{n+1}t^n - \sum_{n=0}^{\infty} \beta^{n+1}t^n.$$

We conclude thus that

$$\sharp E(\mathbb{F}_q) = 1 + q - (\alpha + \beta), \qquad \sharp E(\mathbb{F}_{q^2}) = 1 + q^2 - (\alpha^2 + \beta^2). \tag{12}$$

There are two interesting conclusions from this:

Hasse bound: $\boxed{|\sharp E(\mathbb{F}_q) - (q+1)| \leq 2\sqrt{q}}$

suggesting that our heuristic was spot on. The Hasse bound was known much before the resolution of the Weil conjectures and predates Weil's conjectures. In fact, already when Weil proposed his conjectures, he was able to support them by proving them for all smooth curves, not just elliptic curves.

Another interesting observation is that $\sharp E(\mathbb{F}_q)$ and $\sharp E(\mathbb{F}_{q^2})$ determine ζ_E, because this data allows for solving for the pair α, β.

Example 12.3.1. Let $p = 3$ and consider the elliptic curve $y^2 = x^3 - x$ over $\mathbb{F}_3$. Let's calculate its zeta function. We should remember that we are counting projective solutions and so we always have the point at infinity $0_E = [0 : 1 : 0]$. The squares over $\mathbb{F}_3$ are $0, 1$. Checking one x at the time, we find:

$$E(\mathbb{F}_3) = \{0_E,\ (0,0),\ (1,0),\ (2,0)\}.$$

As 2 is not a square, we can take for the field of 9 elements the following model:

$$\mathbb{F}_9 = \mathbb{F}_3[t]/(t^2 - 2).$$

The squares in $\mathbb{F}_3$ are $(a + b\sqrt{2})^2 = a^2 + 2b^2 + 2ab\sqrt{2}$. Every element in $\mathbb{F}_3$ is a square in $\mathbb{F}_9$. Taking $a = 1$ and $b = 1$ or $b = 2$, we see that also $\sqrt{2}$ and $2\sqrt{2}$ are squares too. Altogether, there are $4 = (9-1)/2$ squares in $\mathbb{F}_9^\times$ and so the squares in $\mathbb{F}_9$ are $\{0, 1, 2, \sqrt{2}, 2\sqrt{2}\}$.

If $x = a + b\sqrt{2}$ then $x^3 = a^3 + 2b^3\sqrt{2} = a + 2b\sqrt{2}$ and so $x^3 - x = b\sqrt{2}$, which is always a square and is zero precisely when $b = 0$. Therefore, $\sharp E(\mathbb{F}_9)$ is equal to 1 (the point at infinity) + 6×2 (coming from all the points $x = a + b\sqrt{2}$ with $b \neq 0$) + 3 (coming from the points $x = a$). Altogether,

$$\sharp E(\mathbb{F}_9) = 16.$$

This still satisfies the Hasse bound for $q = 9$, but just barely.

From the two calculations we conclude that

$$\alpha + \beta = 0, \quad \alpha^2 + \beta^2 = -6.$$

Thus, without loss of generality, $\alpha = \sqrt{-3}, \beta = -\sqrt{-3}$. Therefore,

$$\zeta_E = \frac{1 + 3T^2}{(1 - T)(1 - 3T)}.$$

We can now easily find the number of points of E over any finite field or characteristic 3. For example,

$$\sharp E(\mathbb{F}_{27}) = 1 + 27 - (\alpha^3 + \beta^3) = 28, \qquad \sharp E(\mathbb{F}_{3^4}) = 1 + 81 - (\alpha^4 + \beta^4) = 64.$$

Remark 12.3.2. Here is an important remark. It turns out that for elliptic curves one has more information about the roots of ζ_E. Namely, for elliptic curves over $\mathbb{F}_q$ we also have that

$$\boxed{\alpha\beta = q}$$

This allows us then to find ζ_E *just by calculating* $\sharp E(\mathbb{F}_q)$*!* For the example above, we find $\alpha + \beta = 0$ and $\alpha\beta = 3$, and from this we find that $\alpha = \sqrt{-3}, \beta = -\sqrt{-3}$.

Exercise 12.3.3. Find the number of projective points of the elliptic curve $y^2 = x^3 - 1$ over $\mathbb{F}_5$. Use it to calculate the cardinalities of $\mathbb{E}(\mathbb{F}_{5^2}), \mathbb{E}(\mathbb{F}_{5^3}), \mathbb{E}(\mathbb{F}_{5^4})$; write the zeta function of E as a ratio of explicit polynomials.

12.3.1. *The group law on an elliptic curve.* Let $y^2 = x^3 + ax + b, a, b \in \mathbb{F}$ be an elliptic curve over a field $\mathbb{F}$, finite or infinite. We assume, thus, that this is a non-singular curve. This is the case if and only if the characteristic of $\mathbb{F}$ is not 2 and the polynomial $f(x) = x^3 + ax + b$ is separable (has 3 distinct roots in an algebraic closure $\overline{\mathbb{F}}$). This is not the most general form of an elliptic curve, as one can also consider equations of the form $y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6, a_i \in \mathbb{F}$ (which is necessary for fields of characteristic 2), but it will suffice for our purposes. It is better to consider the projective version

$$E: \quad y^2z = x^3 + axz^2 + bz^3,$$

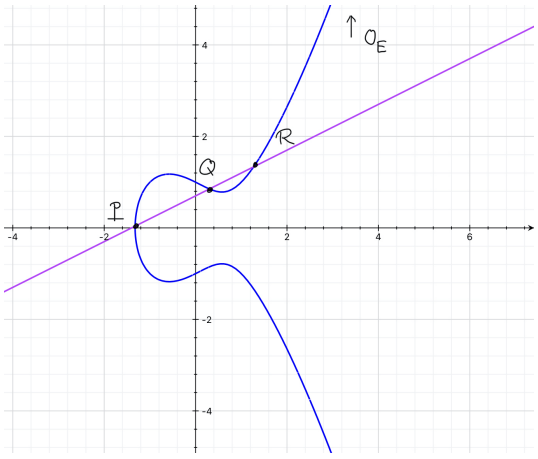
in which just a single point is added: $[x : y : z] = [0 : 1 : 0]$. Denote this point 0_E.

One reason elliptic curves form such an important part of number theory is that there is a group law on elliptic curves, making the solutions $E(\mathbb{L})$, for any field $\mathbb{L} \supseteq \mathbb{F}$ into an abelian group. Of course, if $\mathbb{L}$ is a finite field $E(\mathbb{L})$ is a finite abelian group, but in general $E(\mathbb{L})$ is often infinite. A celebrated theorem states the following:

Theorem 12.3.4 (Mordell-Weil)**.** *Let $\mathbb{F}$ be a finite field extension of the rational numbers $\mathbb{Q}$. Then $E(\mathbb{F})$ is a finitely generated abelian group.*

There is a huge volume of literature concerning the possible structure and rank of the groups $E(\mathbb{F})$, especially for $\mathbb{F} \supseteq \mathbb{Q}$ and for $\mathbb{F}$ a finite field, and many open problems, one of which is the question whether for every integer N there is an elliptic curve E over $\mathbb{Q}$ such that $E(\mathbb{Q})$ has rank at least N. The largest known N is currently (February 2021) $N = 28$, a result due to N. Elkies.

The structure of abelian group on E can be explained as follows (the proof that it is a group law is not easy, especially the associative property, and we will not discuss it here[23]): Since E is a cubic curve in $\mathbb{P}^2$, by Bezout's theorem, any line in $\mathbb{P}^2$ intersects E in 3 points. The group law is defined in such a way that 3 points P, Q, R with coordinates in $\mathbb{F}$, lying on the same line, sum up to 0_E in $E(\mathbb{F})$. The point 0_E is the zero point for this group law, and if $P = (x, y)$ then $-P = (x, -y)$. (The formula for the inverse is more complicated if we work with the more general equation for an elliptic curve.) To calculate $P + P$ use the tangent line at P; the third point of intersection $R = (x, y)$ is $-(P + P)$ and $(x, -y) = P + P$.



[23]The book by Joseph H. Silverman, *The arithmetic of elliptic curves*. Graduate Texts in Mathematics, 106, is an excellent and canonical reference.

In Example 12.3.1, we see that $E(\mathbb{F}_3)$ consists of 4 points that are each equal to their inverses. Thus, as a group $E(\mathbb{F}_3) \cong (\mathbb{Z}/2\mathbb{Z})^2$. It is more laborious, but one can prove by hand that $E(\mathbb{F}_9) \cong (\mathbb{Z}/4\mathbb{Z})^2$. To get more examples, it is better to use PARI. For example, one finds $E(\mathbb{F}_{27}) \cong \mathbb{Z}/14\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$, $E(\mathbb{F}_{81}) \cong (\mathbb{Z}/8\mathbb{Z})^2$.

The group structure of $E(\mathbb{F})$, when $\mathbb{F}$ is a finite field is important to elliptic curves cryptography. In those applications one would like $E(\mathbb{F})$ to be a cyclic group, or almost a cyclic group. For example, one of the curves recommended by NIST in 2013, is a certain elliptic curve E over the field $\mathbb{F}_p$ with

$$p = 2^{192} - 2^{64} - 1 = 6277101735386680763835789423207666416083908700390324961279$$

such that $E(\mathbb{F}_p)$ is a cyclic group of prime order

$$6277101735386680763835789423176059013767194773182842284081.$$

12.4. **A few words about the Sato-Tate conjecture.** Let E be an elliptic curve over $\mathbb{Q}$, say $y^2 = x^3 + Ax + B$, and assume, for simplicity, that $A, B \in \mathbb{Z}$. For a prime p, denote by A_p, B_p the reduction of A, B modulo p. For all primes, but finitely many, we have that the reduction of E modulo p

$$E_p : \quad y^2 = x^3 + A_p x + B_p,$$

is an elliptic curve. Namely, it stays a non-singular equation. Consider the quantity

$$err(p) := \frac{\sharp E(\mathbb{F}) - (p+1)}{2\sqrt{p}}.$$

This is the normalized error relative to the expected number of points, which is $p+1$. Using the Hasse bound we see that

$$err(p) \in [-1, 1].$$

We can therefore write

$$err(p) = \cos(\theta_p), \quad 0 \leq \theta_p \leq \pi.$$

It is natural to ask how this error behaves as the prime p varies. The Sato-Tate conjecture provides an answer to that. To simplify its statement we assume that E does not have "complex multiplication", a property we will not define but remark that it exclude only a "handful" of curves. The Sato-Tate conjecture was proven by Laurent Clozel, Michael Harris, Nicholas Shepherd-Barron, and Richard Taylor in 2008. In fact, they proved a much more general result, but we will not cite it here.

The Sato-Tate conjecture. The statistic of θ_p is given by the following distribution. Let $0 \leq \alpha \leq \beta \leq \pi$. Then,

$$\lim_{N\to\infty} \frac{\#\{p \leq N : \alpha \leq \theta_p \leq \beta\}}{\#\{p \leq N\}} = \frac{2}{\pi} \int_\alpha^\beta \sin^2\theta \, d\theta.$$

Exercise 12.4.1. What is the probability that $|E_p(\mathbb{F}_p) - (p+1)| \leq \sqrt{p}$?

There is another statistics, "orthogonal" to the above, that one may consider. This is the question known as "Sato-Tate on average" and it asks, for a fixed prime p, how does θ_p behave when we vary A_p, B_p. Both the generalization of the Sato-Tate conjecture and its averaged version are subjects of ongoing research.

13. GAUSS SUMS

Gauss sums are examples of so-called trigonometric, or exponential, sums. They were introduced by Gauss who made good use of them to prove the law of quadratic reciprocity, to be discussed later on, and to study cyclotomic fields, which are fields of the form $\mathbb{Q}(e^{2\pi i/n})$ obtained from $\mathbb{Q}$ by adjoining an n-th complex root of unity. There are many more kinds of trigonometric sums and they make an appearance in many branches of number theory, for example in counting the number of points of varieties over finite fields, and so their study is important. In fact, one of the applications given by A. Weil and P. Deligne of the Weil conjectures is to estimate trigonometric sums. A simple example appears in Exercise 76.

Let $\mathbb{F}$ be a finite field, $\mathbb{F} = \mathbb{F}_q, q = p^s$, p prime. Recall the trace map $\mathrm{Tr}_{\mathbb{F}/\mathbb{F}_p}$ that we will simply denote Tr, unless confusion is possible. It is a surjective additive map,

$$\mathrm{Tr}\colon \mathbb{F} \to \mathbb{F}_p.$$

Similarly, we let Nm denote the norm map $\mathrm{Nm}_{\mathbb{F}/\mathbb{F}_p}$, which is a surjective homomorphism,

$$\mathrm{Nm}\colon \mathbb{F}^\times \to \mathbb{F}_p^\times.$$

We remark that there is a unique isomorphism between $\mathbb{F}_p$ to $\mathbb{Z}/p\mathbb{Z}$, determined by $1_{\mathbb{F}} \mapsto 1 \pmod p$ and thus, for all practical purposes, we may think of $\mathbb{F}_p$ as $\mathbb{Z}/p\mathbb{Z}$.

13.1. **Characters.** Let $\zeta_p = e^{2\pi i/p}$; it is a p-th complex root of unity. Define ψ as

$$\psi\colon \mathbb{F} \longrightarrow \mathbb{C}, \qquad \psi(x) = \zeta_p^{\mathrm{Tr}(x)}.$$

If $\mathbb{F} = \mathbb{Z}/p\mathbb{Z}$, then ψ is just the function

$$a \mapsto e^{2\pi i a/p}.$$

The function ψ will be used throughout this chapter. It is sometimes referred to as an **additive character** and should not be confused with the multiplicative characters that we discuss below (those we will mostly denote by χ). Note that ψ mixes "apples and oranges": ζ_p is a complex number and $\mathrm{Tr}(x)$ is an element of the finite field $\mathbb{F}_p$. But, as said, we can identify $\mathrm{Tr}(x)$ unambiguously with a residue class mod p, and since $\zeta^p = 1$, the value $\zeta_p^{\mathrm{Tr}(x)}$ is well-defined. It doesn't matter which integer representative we took for the congruence class mod p.

Lemma 13.1.1. *The function ψ has the following properties:*

(1) $\psi(\alpha+\beta) = \psi(\alpha)\psi(\beta)$. That is, ψ is a group homomorphism from $\mathbb{F}$, viewed as a group under addition, and $\mathbb{C}^\times$, viewed as a group under multiplication. In fact, it has image in

$$\mu_p := \{z \in \mathbb{C} : z^p = 1\},$$

which is a cyclic group of order p generated by ζ_p.

(2) $\exists \alpha \in \mathbb{F}$ such that $\psi(\alpha) \neq 1$ and so ψ is surjective onto μ_p.

(3) $\sum_{\alpha\in\mathbb{F}} \psi(\alpha) = 0$.

Proof. (1) follows immediately from the additivity of Tr, and (2) from the fact that Tr is surjective. To prove (3), choose β such that $\psi(\beta) \neq 1$. Then,

$$\psi(\beta)\sum_{\alpha\in\mathbb{F}} \psi(\alpha) = \sum_{\alpha\in\mathbb{F}} \psi(\alpha+\beta) = \sum_{\alpha\in\mathbb{F}} \psi(\alpha),$$

because when α varies over $\mathbb{F}$ so does $\alpha+\beta$. Thus, $\sum_{\alpha\in\mathbb{F}} \psi(\alpha) = 0$.[24] □

[24] Variants of this trick for proving that a certain sum is equal to 0 will be used over and over below.

Let $\delta_{x,y}$ denote the Kronecker δ-function. Namely, $\delta_{x,y} = 1$ if $x = y$ and $\delta_{x,y} = 0$ if $x \neq y$.

Corollary 13.1.2. *Let $x, y \in \mathbb{F}$. We have*

$$\frac{1}{q}\sum_{\alpha\in\mathbb{F}} \psi(\alpha(x-y)) = \delta_{x,y}.$$

Proof. When $x = y$ this follows from $\psi(0) = 1$. When $x \neq y$, $\alpha(x-y)$ ranges over all elements of $\mathbb{F}$ as α ranges over $\mathbb{F}$, and so $\sum_{\alpha\in\mathbb{F}} \psi(\alpha(x-y)) = \sum_{\alpha\in\mathbb{F}} \psi(\alpha) = 0$. □

By a **character** χ of $\mathbb{F}$ we mean a group homomorphism:

$$\chi\colon \mathbb{F}^\times \to \mathbb{C}^\times, \quad \chi(xy) = \chi(x)\chi(y).$$

The simplest example is provided by the **trivial character**,

$$\epsilon\colon \mathbb{F}^\times \to \mathbb{C}^\times, \quad \epsilon(x) = 1, \forall x \in \mathbb{F}^\times.$$

We extend χ to $\mathbb{F}$ as follows:

$$\chi(0) = \begin{cases} 0, & \chi \neq \epsilon; \\ 1, & \chi = \epsilon. \end{cases}$$

Note that although $\mathbb{F}$ is not a group under multiplication, the identity $\chi(xy) = \chi(x)\chi(y)$ still holds for all $x, y \in \mathbb{F}$.

Example 13.1.3. Assume that $\mathbb{F} = \mathbb{F}_p$. Define the **Legendre symbol** $\left(\frac{\cdot}{p}\right)$: Let a be a congruence class modulo p.

$$\left(\frac{a}{p}\right) = \begin{cases} 1, & a \text{ is a square, and } a \neq 0; \\ -1, & a \text{ is not a square, and } a \neq 0; \\ 0, & a = 0. \end{cases}$$

Sometimes, to improve type setting, we will want a more compact notation. We will then use λ to denote this character:

$$\lambda(a) := \left(\frac{a}{p}\right).$$

The behaviour of this function as a ranges from 1 to p is rather hard to predict. Here is a table of the some values of $\left(\frac{a}{p}\right)$.

prime	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
3	1	-1																				
5	1	-1	-1	1																		
7	1	1	-1	1	-1	-1																
11	1	-1	1	1	1	-1	-1	-1	1	-1												
13	1	-1	1	1	-1	-1	-1	-1	1	1	-1	1										
17	1	1	-1	1	-1	-1	-1	1	1	-1	-1	-1	1	-1	1	1						
19	1	-1	-1	1	1	1	1	-1	1	-1	1	-1	-1	-1	-1	1	1	-1				
23	1	1	1	1	-1	1	-1	1	1	-1	-1	1	1	-1	-1	1	-1	1	-1	-1	-1	-1

TABLE 1. The Legendre symbol

Suppose now that $p \neq 2$. Then, the homomorphism $\mathbb{F}^\times \to \mathbb{F}^\times, x \mapsto x^2$, has kernel $\{\pm 1\}$, hence the image, the squares in $\mathbb{F}^\times$ are a subgroup of $\mathbb{F}^\times$ with $(p-1)/2$ elements. Let us denote it $\mathbb{F}^{\times,2}$. It follows that $\mathbb{F}^\times/\mathbb{F}^{\times,2}$ is a group with two elements, and the Legendre symbol is an isomorphism

$$\mathbb{F}^\times/\mathbb{F}^{\times,2} \xrightarrow{\sim} \{\pm 1\}\ , \qquad a \mapsto \left(\frac{a}{p}\right).$$

In particular, the Legendre symbol is a character. This has interesting applications. For example, it implies that the product of two non-zero non-squares is always a square, and the same for their ratio.

Lemma 13.1.4. *Let χ be a character of $\mathbb{F}^\times$.*

(1) If $\chi \neq \epsilon$ then $\sum_{t\in\mathbb{F}} \chi(t) = 0$.
(2) If $\chi = \epsilon$ then $\sum_{t\in\mathbb{F}} \chi(t) = q$.

Exercise 13.1.5. Prove Lemma 13.1.4. (Compare with the proof of Lemma 13.1.1.)

Suppose that χ_1, χ_2 are characters. Then $\chi_1\chi_2$ is a character too. Indeed, for $x, y \in \mathbb{F}^\times$ we have, by definition of $\chi_1\chi_2$,

$$(\chi_1\chi_2)(xy) = \chi_1(xy)\chi_2(xy) = \chi_1(x)\chi_1(y)\chi_2(x)\chi_2(y) = (\chi_1\chi_2)(x)\cdot(\chi_1\chi_2)(y).$$

However, when we view χ_1, χ_2 as extended to $\mathbb{F}$, some care is needed. If $\chi_1\chi_2 = \epsilon$ and, say, $\chi_1 \neq \epsilon$, then $(\chi_1\chi_2)(0) = \epsilon(0) = 1 \neq \chi_1(0)\chi_2(0)$. That is, the extension of $\chi_1\chi_2$ to $\mathbb{F}$ is done as the extension of the character $\chi_1\chi_2$ on $\mathbb{F}^\times$ to $\mathbb{F}$, and its value at 0 may be different than $\chi_1(0)\chi_2(0)$.

We also note that if we let χ^{-1} be defined on $\mathbb{F}^\times$ by $\chi^{-1}(a) = \chi(a)^{-1}$ then we have $\chi\chi^{-1} = \epsilon$ and

$$\chi^{-1}(a) = \chi(a)^{-1} = \overline{\chi(a)},$$

because any character on $\mathbb{F}^\times$ takes values in roots of unity. Because of this identity, we also use the notation $\bar{\chi}$ for the inverse:

$$\bar{\chi} = \chi^{-1}.$$

Proposition 13.1.6. *The characters of $\mathbb{F}^\times$ form a group under multiplication. This group is naturally isomorphic to the multiplicative group μ_{q-1} of complex roots of unity of order $q-1$. In addition, if $a \in \mathbb{F}^\times$, $a \neq 1$, then there a character χ such that $\chi(a) \neq 1$.*

Proof. We have just explained that the characters form a group. Let g be a generator for the group $\mathbb{F}^\times$. Any character χ is uniquely determined by $\chi(g)$. Indeed, any element of $\mathbb{F}^\times$ is of the form g^n for some integer n, and

$$\chi(g^n) = \chi(g)^n.$$

Furthermore, as $g^{q-1} = 1$, we must have $\chi(g) \in \mu_{q-1}$.

Conversely, given a $(q-1)$-st root of unity ζ, define a character χ by

$$\chi(g^n) = \zeta^n.$$

This shows that we have a bijection between characters and elements of μ_{q-1}:

$$\chi \mapsto \chi(g).$$

By definition of the product of characters this bijection is an isomorphism of groups.

Given an element $a \neq 1$ of $\mathbb{F}$, we can write $a = g^n$ for some $1 \leq n < q-1$. Taking the character corresponding to $\zeta = \exp(2\pi i/(q-1))$, we have $\chi(a) = \zeta^n = \exp(2n\pi i/(q-1)) \neq 1$. □

Let us denote the group of characters of $\mathbb{F}^\times = \mathbb{F}_q^\times$ by $\mathbb{X}_q$. It is a group under multiplication, where $(\chi_1\chi_2)(a) := \chi_1(a)\chi_2(a)$ for $a \in \mathbb{F}^\times$. The identity element is ϵ. Once we have chosen a generator g for $\mathbb{F}^\times$, we have an isomorphism

$$\mathbb{X}_q \cong \mu_{q-1}.$$

It will also be useful to denote by χ_ζ the character corresponding to $\zeta \in \mu_{q-1}$ under this isomorphism;

$$\chi_\zeta(g^n) = \zeta^n.$$

For every integer n we denote by $\mathbb{X}_q[n]$ the elements of order dividing n in $\mathbb{X}_q$:

$$\mathbb{X}_q[n] = \{\chi \in \mathbb{X}_q : \chi^n = \epsilon\}.$$

This is a group whose order is $\gcd(n, q-1)$; in particular it has order n if $n|(q-1)$. Indeed, in this case

$$\mathbb{X}_q[n] = \{\chi_\zeta : \zeta \in \mu_n\}.$$

We now have the lemma dual to Lemma 13.1.4.

Lemma 13.1.7. *Let $a \in \mathbb{F}^\times$. Then:*

(1) If $a \neq 1$ then $\sum_{\chi\in\mathbb{X}_q} \chi(a) = 0$.
(2) If $a = 1$ then $\sum_{\chi\in\mathbb{X}_q} \chi(a) = q-1$.

Proof. The second statement is clear. For the first we use a trick we have seen before. Choose some χ_0 such that $\chi_0(a) \neq 1$; such χ_0 exists by Proposition 13.1.6. Then,

$$\chi_0(a)\sum_{\chi\in\mathbb{X}_q} \chi(a) = \sum_{\chi\in\mathbb{X}_q} (\chi_0\chi)(a) = \sum_{\chi\in\mathbb{X}_q} \chi(a),$$

and this implies $\sum_{\chi\in\mathbb{X}_q} \chi(a) = 0$. We have used the fact that if G is a group, and g is any element of G, then $\{gx : x \in G\} = G$ and applied it for $G = \mathbb{X}_q, g = \chi_0$. □

13.2. **The equation $x^n = a$.** Our purpose in this section is to show that there is a connection between characters and counting the number of solutions to equations over $\mathbb{F}$. Although the case we consider here is very simple, the idea is absolutely fundamental to everything that follows.

Proposition 13.2.1. *Let $n|(q-1)$ be a positive integer and let $a \in \mathbb{F} = \mathbb{F}_q$. Denote by*

$$N(x^n = a)$$

the number of solutions to the equation $x^n = a$ in $\mathbb{F}$. Then,

$$\boxed{N(x^n = a) = \sum_{\chi\in\mathbb{X}_q[n]} \chi(a)}.$$

Proof. If $a = 0$, $N(x^n = a) = 1$. On the other hand, for every $\chi \neq \epsilon$, $\chi(0) = 0$. Therefore,

$$\sum_{\chi\in\mathbb{X}_q[n]} \chi(0) = \epsilon(0) = 1 = N(x^n = a),$$

and we get the equality we want.

Suppose then that $a \neq 0$. Note that $\mu_n(\mathbb{F})$, the group of n-th root of unity in $\mathbb{F}$,

$$\mu_n(\mathbb{F}) := \{u \in \mathbb{F}^\times : u^n = 1\},$$

is a cyclic group of order n, because $n|(q-1)$ and $\mathbb{F}^\times$ is a cyclic group of order $q-1$.

If a is not an n-th power in $\mathbb{F}$ then $N(x^n = a) = 0$. On the other hand, if a is an n-th power in $\mathbb{F}$ then $N(x^n = a) = n$, because if $b^n = a$ is one solution then all other solutions are of the form $(bu)^n = a$, where $u \in \mu_n(\mathbb{F})$ and there are precisely n of them. We need to show that $\sum_{\chi \in \mathbb{X}_q[n]} \chi(a)$ has the same behaviour.

In the second case, $a = b^n$, $a \neq 0$, we have

$$\sum_{\chi \in \mathbb{X}_q[n]} \chi(a) = \sum_{\chi \in \mathbb{X}_q[n]} \chi(b^n) = \sum_{\chi \in \mathbb{X}_q[n]} \chi^n(b) = \sum_{\chi \in \mathbb{X}_q[n]} \epsilon(b) = n.$$

Consider now the first case: $a \neq b^n$ for any b. Consider χ_ζ, where $\zeta = \exp(2\pi i/n)$. Suppose that $\chi_\zeta(a) = 1$. Let us write $a = g^r$ for some r. Then

$$\chi_\zeta(a) = \zeta^r = 1.$$

But this implies $n|r$ and so that a is an n-th power, which is not the case. Therefore,

$$\chi_\zeta(a) \neq 1.$$

Note that $\chi_\zeta^n = \chi_{\zeta^n} = \chi_1 = \epsilon$. That is, $\chi_\zeta \in \mathbb{X}_q[n]$. We can now perform what is by-now an old trick:

$$\chi_\zeta(a) \sum_{\chi \in \mathbb{X}_q[n]} \chi(a) = \sum_{\chi \in \mathbb{X}_q[n]} (\chi_\zeta \chi)(a) = \sum_{\chi \in \mathbb{X}_q[n]} \chi(a),$$

and we conclude that $\sum_{\chi \in \mathbb{X}_q[n]} \chi(a) = 0$. □

Example 13.2.2. Suppose that p is odd. Then $\mathbb{X}_q$, being cyclic of even order $q-1$ has only one element of order 2. Since

$$a \mapsto \left(\frac{\mathrm{Nm}(a)}{p}\right),$$

cannot be the trivial character (as Nm is surjective and p is odd), it is a character of order 2 and we find that

$$\mathbb{X}_q[2] = \left\{\epsilon, \left(\frac{\mathrm{Nm}(\cdot)}{p}\right)\right\}.$$

Consequently,

$$N(x^2 = a) = 1 + \left(\frac{\mathrm{Nm}(a)}{p}\right).$$

Now, if $\mathbb{F} = \mathbb{F}_p$, this is clear from the definition of the Legendre symbol. But, in general some thought is required to see directly why this is true.

Exercise 13.2.3. Assume that p is an odd prime, $q = p^s$. Prove the formula $N(x^2 = a) = 1 + \left(\frac{\mathrm{Nm}(a)}{p}\right)$ for $\mathbb{F}_q$ by proving that a is a square in $\mathbb{F}_q$ if and only if $\mathrm{Nm}(a)$ is a square in $\mathbb{F}_p$.

13.3. **Definition and first properties of Gauss sums.** Let $\mathbb{F} = \mathbb{F}_q, q = p^s$, be a finite field with q elements of characteristic p. Let ψ be as in §13.1. Let $\chi \in \mathbb{X}_q$ be a character of $\mathbb{F}^\times$ and let $a \in \mathbb{F}$. The **Gauss sum** associated to χ and a is defined as

$$\mathfrak{g}_a(\chi) = \sum_{t \in \mathbb{F}} \chi(t)\psi(at).$$

The special case of $a = 1$ is the most important. In this case we just use the notation $\mathfrak{g}(\chi)$. Thus,

$$\mathfrak{g}(\chi) = \sum_{t \in \mathbb{F}} \chi(t)\psi(t).$$

We remark that $\mathfrak{g}_a(\chi)$ is a complex number, which is a sum of roots of unity, but not a root of unity itself; in fact, we shall prove that if $\chi \neq \epsilon$ its absolute value is $\sqrt{q}$.

Example 13.3.1. A good case to keep in mind is when $\mathbb{F} = \mathbb{F}_p$. Then,

$$\mathfrak{g}(\chi) = \sum_{t=0}^{p-1} \chi(t) e^{\frac{2\pi i}{p} \cdot t}, \qquad \mathfrak{g}_a(\chi) = \sum_{t=0}^{p-1} \chi(t) e^{\frac{2\pi i}{p} \cdot at}.$$

The next proposition explains why $\mathfrak{g}(\chi)$ is the central definition.

Proposition 13.3.2. *The Gauss sums have the following properties:*

$$\mathfrak{g}_a(\chi) = \begin{cases} \chi(a^{-1}) \cdot \mathfrak{g}(\chi) & a \neq 0, \chi \neq \epsilon, \\ q & a = 0, \chi = \epsilon, \\ 0 & a = 0, \chi \neq \epsilon, \\ 0 & a \neq 0, \chi = \epsilon. \end{cases}$$

Proof. If $a \neq 0$ we have

$$\mathfrak{g}(\chi) = \sum_{t \in \mathbb{F}} \chi(t)\psi(t) = \sum_{t \in \mathbb{F}} \chi(at)\psi(at) = \chi(a) \sum_{t \in \mathbb{F}} \chi(t)\psi(at) = \chi(a)\mathfrak{g}_a(\chi).$$

Thus, if $a \neq 0$, $\mathfrak{g}_a(\chi) = \chi(a^{-1})\mathfrak{g}(\chi)$. This prove the first case. To prove the last case we need to show that if $a \neq 0$ and $\chi = \epsilon$ then

$$\mathfrak{g}_a(\epsilon) = \sum_{t \in F} \psi(at) = \sum_{t \in F} \psi(t) = 0,$$

but this is Lemma 13.1.1 (3). As $\psi(0) = 1$, the two remaining cases follow directly from Lemma 13.1.4. □

Theorem 13.3.3. *If $\chi \neq \epsilon$,*

$$|\mathfrak{g}_a(\chi)| = \sqrt{q}.$$

Proof. As $a \neq 0$, $\chi(a)$ is a root of unity and so, using Proposition 13.3.2, it is enough to prove $|\mathfrak{g}(\chi)| = \sqrt{q}$. The proof is based on evaluating $A := \sum_{a \in \mathbb{F}} \mathfrak{g}_a(\chi)\overline{\mathfrak{g}_a(\chi)}$ in two different ways.

On the one hand, using that $\bar{\chi} = \chi^{-1}$ and Proposition 13.3.2,

$$A = \sum_{a \in \mathbb{F}} \mathfrak{g}_a(\chi)\overline{\mathfrak{g}_a(\chi)} = \sum_{a \in \mathbb{F}^\times} \mathfrak{g}(\chi)\overline{\mathfrak{g}(\chi)}\chi(a^{-1})\bar{\chi}(a^{-1}) = (q-1)|\mathfrak{g}(\chi)|^2.$$

Note that $\overline{\psi(at)} = \zeta_p^{-\mathrm{Tr}(at)} = \zeta_p^{\mathrm{Tr}(-at)}$. Thus, on the other hand, from the definition of $\mathfrak{g}_a(\chi)$, we have

$$\overline{\mathfrak{g}_a(\chi)} = \sum_{t \in \mathbb{F}} \overline{\chi(t)\psi(at)} = \sum_{t \in \mathbb{F}} \overline{\chi(t)}\psi(-at).$$

Thus, using Corollary 13.1.2,

$$A = \sum_{x,y \in \mathbb{F}} \sum_{a \in \mathbb{F}} \chi(x)\bar{\chi}(y)\psi(ax - ay) = \sum_{x,y \in \mathbb{F}^\times} \chi(xy^{-1})q\delta_{xy} = (q-1)q.$$

Comparing the two expression for A, the proof is complete. □

Corollary 13.3.4. *We have*

$$\overline{\mathfrak{g}(\chi)} = \chi(-1)\mathfrak{g}(\bar{\chi}), \quad \textit{and} \quad \mathfrak{g}(\chi) \cdot \mathfrak{g}(\bar{\chi}) = \chi(-1) \cdot q.$$

Proof. We know that $\mathfrak{g}(\chi)\overline{\mathfrak{g}(\chi)} = q$. But,

$$\overline{\mathfrak{g}(\chi)} = \sum_{t \in \mathbb{F}} \overline{\chi(t)\psi(t)} = \bar{\chi}(-1) \sum_{t \in \mathbb{F}} \bar{\chi}(-t)\psi(-t) = \bar{\chi}(-1)\mathfrak{g}(\bar{\chi}) = \chi(-1)\mathfrak{g}(\bar{\chi}),$$

where we used that $\chi(-1) \in \{\pm 1\}$ so $\bar{\chi}(-1) = \chi(-1)$. The Corollary follows. □

Remark 13.3.5. View $\bar{\mathbb{Q}}$ as a subset of $\mathbb{C}$. The formula $\overline{\mathfrak{g}(\chi)} = \chi(-1)\mathfrak{g}(\bar{\chi})$ is a special case of an action of an automorphism of $\bar{\mathbb{Q}}$, viz. complex conjugation, on Gauss sums. A reader familiar with Galois theory should work out the general case.

Theorem 13.3.3 shows that the Gauss sum, which is a sum of q roots of unity, and so could *a priori* have absolute value as large as q, actually has absolute value $\sqrt{q}$ (unless $\chi = \epsilon$). There is a lot of cancellation going on. Consider the case when $q = p$ and χ is the Legendre symbol. Then the theorem says that

$$\Big| \sum_{a=0}^{p-1} \left(\frac{a}{p}\right) e^{\frac{2\pi i a}{p}} \Big| = \sqrt{p}.$$

On the one hand, this supports our statement that the Legendre symbol $\left(\frac{a}{p}\right)$ behaves "erratically" as a varies (and cf. Table 13.1.3). On the other hand, if the behaviour was *truly* random, one would not expect the Gauss sum to have absolute value $\sqrt{p}$ *on the nose*, but rather to have absolute value *about* $\sqrt{p}$, with the exact value $\sqrt{p}$ occurring very rarely.

Another remarkable evidence to the random behaviour of $\left(\frac{a}{p}\right)$ is the Polya-Vinogradov inequality that we state here only for the Legendre symbol, although it holds in much greater generality.

Theorem 13.3.6 (Polya-Vingoradov)**.** *For any integers $m \leq n$,*

$$\Big| \sum_{a=m}^{n} \left(\frac{a}{p}\right) \Big| < \sqrt{p} \log p.$$

Proof. For the proof, it is more elegant to change notation and work with a sum from m to $n-1$. Let λ denote the Legendre symbol. Multiplying the left hand side by $|\mathfrak{g}(\lambda)| = \sqrt{p}$ we need to show that

$$\Big| \sum_{a=m}^{n-1} \lambda(a)\mathfrak{g}(\lambda) \Big| < p \log p.$$

Using that $\lambda(a) = \lambda(a^{-1})$ for $a \neq 0$,[25] we get $\lambda(a)\mathfrak{g}(\lambda) = \mathfrak{g}_a(\lambda)$, and this is also true for $a = 0$. We thus need to study the sum $\sum_{a=m}^{n-1} \mathfrak{g}_a(\lambda)$. Using the definition of the Gauss sum, we have

$$\sum_{a=m}^{n-1} \mathfrak{g}_a(\lambda) = \sum_{t \in \mathbb{F}_p} \lambda(t) \sum_{a=m}^{n-1} \psi(at) = \sum_{t \in \mathbb{F}_p^\times} \lambda(t) \sum_{a=m}^{n-1} \beta_t^a = \sum_{t \in \mathbb{F}_p^\times} \lambda(t) \cdot \beta_t^m \, \frac{\beta_t^{n-m} - 1}{\beta_t - 1},$$

where we put $\beta_t = e^{2\pi i t/p}$ and summed the geometric series. We next use the following identities (for the second one use $e^{i\theta} = \cos(\theta) + i\sin(\theta)$):

$$e^{2i\theta} - 1 = e^{i\theta}(e^{i\theta} - e^{-i\theta}) = e^{i\theta} \cdot 2i \sin(\theta).$$

As $\lambda(t), \beta_t$ are roots of unity, we obtain,

$$\Big| \sum_{a=m}^{n-1} \mathfrak{g}_a(\lambda) \Big| = \Big| \sum_{t \in \mathbb{F}_p^\times} \lambda(t) \beta_t^{m + \frac{(n-m-1)}{2}} \cdot \frac{\sin(\pi t(n-m)/p)}{\sin(\pi t/p)} \Big| \leq \sum_{t=1}^{p-1} \frac{1}{|\sin(\pi t/p)|}.$$

It remains to estimate the last sum.

[25] The reader will note that, except for here, never do we need to know what λ is, except that it is not trivial. Examining the proof, with λ replaced by a general character $\chi \in \mathbb{X}_p$, shows that if we multiply instead by $\mathfrak{g}(\bar{\chi})$, the proof goes through. Thus, the proof applies to any $\chi \in \mathbb{X}_p^*$.

Let $\langle x \rangle$ denote that difference between x and the closest integer. Namely, it is the minimum of $|x-n|$ as n ranges over $\mathbb{Z}$. Using the periodicity of $|\sin(x)|$ and that it is symmetric about 0, we see that the last sum is equal to

$$\sum_{t=1}^{p-1} \frac{1}{|\sin(\pi\langle t/p\rangle)|} = 2 \sum_{t=1}^{(p-1)/2} \frac{1}{|\sin(\pi\langle t/p\rangle)|}.$$

However, for $0 \leq x \leq 1/2$, we have $\sin(\pi x) \geq 2x$ (we leave that as an exercise), and so we have the estimate

$$\sum_{t=1}^{p-1} \frac{1}{|\sin(\pi\langle t/p\rangle)|} \leq 2 \sum_{t=1}^{(p-1)/2} \frac{1}{2t/p} = p \sum_{t=1}^{(p-1)/2} \frac{1}{t}.$$

The only remaining point is to show that

$$\sum_{t=1}^{(p-1)/2} \frac{1}{t} < \log(p).$$

We leave that as an exercise (compare $1/x$ with $\log((2x+1)/(2x-1))$). □

Note that this theorem bounds the number of consecutive a's such that a is quadratic residue mod p (and similarly, for non-quadratic residue). A stronger bound and a stronger inequality were found by D. A. Burgess.[26] Interestingly, Burgess' proof uses the Weil conjectures for curves over finite fields, more specifically the Riemann hypothesis part; see Exercise 76. As we remarked before, Weil was able to prove his conjectures for curves at the time he stated them.

Exercise 13.3.7. Consider the case of $\mathbb{F}_p$ and let $\lambda(a) = \left(\frac{a}{p}\right)$ be the Legendre symbol. Suppose that $p \nmid a$. By considering two ways to evaluate the sum $\sum_{n=0}^{p-1} \left(1 + \left(\frac{n}{p}\right)\right) e^{2\pi i a n/p}$, prove that

$$\mathfrak{g}(\lambda) = \sum_{n=0}^{p-1} e^{\frac{2\pi i a n^2}{p}}.$$

13.4. **Quadratic reciprocity.** The law of quadratic reciprocity was conjectured by L. Euler and A.-M. Legendre, and proven by C. F. Gauss who supplied several proofs for it and referred to it as *Theorema Aureum* ("golden theorem"). There are many proofs known for the law of quadratic reciprocity. Here we will show how the theory of Gauss sums may be used to prove this law. Let us first state it.

Recall that for a prime p we have the Legendre symbol $\left(\frac{a}{p}\right)$, which is equal to 1 if a is a non-zero square in $\mathbb{F}_p$, 0 if $a = 0$, and -1 otherwise. See Example 13.1.3

The law of quadratic reciprocity. Let $p \neq q$ be odd primes. Then

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

The law of quadratic reciprocity has two additional complementary statements: Let p be an odd prime. Then,

- $\left(\frac{-1}{p}\right) = (-1)^{\frac{p-1}{2}}$;
- $\left(\frac{2}{p}\right) = (-1)^{\frac{p^2-1}{8}}$.

[26] D. A. Burgess, The distribution of quadratic residues and non-residues. *Mathematika* 4 (1957), 106–112.

The statement concerning $\left(\frac{2}{p}\right)$ is not that easy. But the first statement is not too hard to prove. More generally, for $a \in \mathbb{F}_q, a \neq 0$, prove that a is a square in $\mathbb{F}_q^\times$ if and only if $a^{\frac{q-1}{2}} = 1$.

The law of quadratic reciprocity is really quite astounding at first sight. For example, if $p \equiv 1 \pmod 4$ then it states that p is a square mod q if and only if q is a square mod p. It is hard to see how a statement about something happening mod q could be related to something happening mod p.

Example 13.4.1. Let us answer the question whether the equation

$$x^2 + 28 \equiv 0 \pmod{113}$$

has a solution. Of course, a solution exists if and only if -28 is a square modulo 113. Namely, since 113 is a prime, if and only if $\left(\frac{-28}{113}\right) = 1$. We have,

$$\left(\frac{-28}{113}\right) = \left(\frac{-1}{113}\right)\left(\frac{4}{113}\right)\left(\frac{7}{113}\right) = (-1)^{(113-1)/2}\left(\frac{7}{113}\right) = \left(\frac{113}{7}\right) = \left(\frac{1}{7}\right) = 1.$$

Thus, a solution exists; in fact, two.

Exercise 13.4.2. How many solutions do the following equations have?

(1) $x^2 + 120 \equiv 0 \pmod{257}$.
(2) $x^2 - x - 1 \equiv 0 \pmod{p}$, where $p > 5$ is a prime.

Exercise 13.4.3. Find a prime $p > 2$ such that $1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11$ are all squares modulo p. You may use a computer for some of the computations.

13.4.1. *Proof of the law of quadratic reciprocity.* Let $p \neq q$ be odd primes and let

$$\lambda(t) = \left(\frac{t}{p}\right)$$

be the Legendre symbol.

Claim: $\lambda(a) = a^{\frac{p-1}{2}} \pmod{p}$.

This is true for $a = 0$. Note that for $a \neq 0$, $(a^{(p-1)/2})^2 = a^{p-1} = 1 \pmod{p}$ and so in any case $(a^{(p-1)/2}) \in \{\pm 1\}$.

For $a \neq 0$, as $\mathbb{F}_p^\times$ is cyclic, the subgroup of squares $\mathbb{F}^{\times,2}$ in $\mathbb{F}^\times$ is a cyclic subgroup as well. It is of order $(p-1)/2$ and so it is equal to all the elements of $\mathbb{F}^\times$ whose order divides $(p-1)/2$. Thus, a is a square modulo p if and only if $a^{(p-1)/2} = 1 \pmod{p}$. This proves the Claim.

Now to the proof of Quadratic Reciprocity. To simplify notation, put

$$\gamma = \mathfrak{g}(\lambda) = \sum_{t=0}^{p-1}\left(\frac{t}{p}\right)\zeta_p^t, \qquad p^* = (-1)^{(p-1)/2}p.$$

The main idea of the proof is to calculate γ^q in two ways. One using Corollary 13.3.4; the other, calculating it mod q, using the definition of γ as a sum and the binomial formula modulo q.

First, apply Corollary 13.3.4 to the character λ and note that since λ is of order 2, $\lambda = \lambda^{-1}$; that is, $\lambda = \bar{\lambda}$. We thus find,

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

The following computations are done in the ring

$$\mathbb{Z}[\zeta_p] = \mathbb{Z} \oplus \mathbb{Z}\zeta_p \oplus \cdots \oplus \mathbb{Z}\zeta_p^{p-1} \subset \mathbb{Q}(\zeta_p) \subset \mathbb{C},$$

modulo the principal ideal $(q) = q\mathbb{Z}[\zeta_p]$.

As $\gamma^2 = p^*$, also $\gamma^2 = p^* \pmod{q}$, and thus we have

$$\gamma^{q-1} = (p^*)^{(q-1)/2} = \left(\frac{p^*}{q}\right) \pmod{q},$$

and this implies

$$\gamma^q = \left(\frac{p^*}{q}\right) \gamma \pmod{q}.$$

On the other hand, still calculating mod q and using the binomial formula mod q, we find

$$\gamma^q = \left(\sum_{t \in \mathbb{F}} \left(\frac{t}{p}\right) \zeta_p^t\right)^q = \sum_{t \in \mathbb{F}} \left(\frac{t}{p}\right) \zeta_p^{qt} = \mathfrak{g}_q(\lambda) = \left(\frac{q^{-1}}{p}\right) \mathfrak{g}(\lambda) = \left(\frac{q}{p}\right) \gamma \mod (q),$$

where we have used that λ is quadratic and so $\lambda(q) = \lambda(q^{-1})$.

Now, note that γ is not a zero-divisor in $\mathbb{Z}[\zeta_p]/(q)$. In fact, as $\gamma^2 = \pm p$ is invertible mod (q), so is γ. Comparing the two expressions for γ^q, we conclude that

$$\left(\frac{p^*}{q}\right) = \left(\frac{(-1)^{(p-1)/2} p}{q}\right) = \left(\frac{(-1)^{(p-1)/2}}{q}\right) \left(\frac{p}{q}\right) = (-1)^{\frac{(p-1)(q-1)}{4}} \left(\frac{p}{q}\right) = \left(\frac{q}{p}\right) \mod (q).$$

But this is a mod q congruence of integers in $\{\pm 1\}$ and $q > 2$. So we can deduce that as integers

$$(-1)^{\frac{(p-1)(q-1)}{4}} \left(\frac{p}{q}\right) = \left(\frac{q}{p}\right),$$

which is the law of quadratic reciprocity.

□

13.4.2. *An application of quadratic reciprocity.* The **Fermat numbers** are integers of the form

$$F_n = 2^{2^n} + 1, \quad n \in \mathbb{N}.$$

If they are prime they are called **Fermat primes**. Fermat primes are interesting for example because of the problem of constructing regular polygons in the plane using only a compass and straightedge; a problem that dates back to ancient Greece. A theorem usually proved in a first course about fields is that a regular n-gon can be thus constructed if and only if

$$n = 2^k p_1 p_2 \dots p_r,$$

where the p_i are distinct Fermat primes and k any non-negative integer. The first Fermat numbers $F_0, \dots, F_4$ are

$$3, \ 5, \ 17, \ 257, \ 65537,$$

and are primes. However,

$$F_5 = 641 \times 6700417.$$

At this point (February 2021), there are no more known examples of Fermat primes. P. de Fermat thought that F_n are always primes, but it is known that $F_5, F_6, \dots, F_{32}$ are composite. The largest Fermat number known to be composite (October 2020) is $F_{18233954}$; this was discovered by Ryan

Propper.[27] In fact, the experts suspect that there are no Fermat primes, besides those listed above.

One might wonder why one takes the special exponent 2^{2^n} and not just 2^n. The answer is provided by the following Proposition.

Proposition 13.4.4. *Assume that $2^m + 1$ is prime. Then $m = 2^n$ for some n.*

Proof. Suppose that $m = pM$, with p an odd prime. Then, from the factorization $x^p + 1 = (x+1)(x^{p-1} - x^{p-2} + \cdots + 1)$, we find

$$(2^m + 1) = (2^M)^p + 1 = (2^M + 1)(2^{(p-1)M} - 2^{(p-2)M} + \cdots + 1),$$

and therefore $2^m + 1$ is composite. □

Our purpose now is to explain how one tests whether Fermat numbers are prime. With start with the Lucas-Lehmer test for primality

Theorem 13.4.5 (Lucas-Lehmer)**.** *Let n be a positive integer. Suppose that there is a positive integer a such that $a^{n-1} \equiv 1 \pmod{n}$, but for every prime divisor p of $n-1$ we have $a^{\frac{n-1}{p}} \not\equiv 1 \pmod{n}$. Then, n is prime.*

Proof. We work in the group $(\mathbb{Z}/n\mathbb{Z})^\times$, which is the group of units of the ring $\mathbb{Z}/n\mathbb{Z}$; it is a group under multiplication and its order is, by definition, $\varphi(n)$. Here φ is Euler's function:

$$\varphi(n) = \sharp\{1 \le i \le n : \gcd(i, n) = 1\}.$$

It is clear from the definition that $\varphi(n) = n - 1$ if and only if n is prime.

Let a be as in the statement of the theorem. If $d|a$ then $d|a^{n-1} = 1 + kn$, for some k. It follows that if $d > 1$, $d \nmid n$. That is, the condition on a implies that $a \in (\mathbb{Z}/n\mathbb{Z})^\times$.

We claim that the order of a is precisely $n - 1$. If not, the order of a is some integer m where m is a proper divisor of $n - 1$ and thus $m|\frac{n-1}{p}$ for some prime p. It then follows that $a^{\frac{n-1}{p}} = (a^m)^{\frac{n-1}{mp}} = 1$. Contradiction.

Since the order of an element divides the order of the group, and since the order of $(\mathbb{Z}/n\mathbb{Z})^\times$ is $\varphi(n)$, we conclude that $(n-1)|\varphi(n)$. This implies that $n - 1 = \varphi(n)$ and so that n is prime. □

Theorem 13.4.6 (T. Pépin's test)**.** *The Fermat number F_n is prime if and only if*

$$3^{(F_n-1)/2} \equiv -1 \pmod{F_n}.$$

Proof.

Exercise 13.4.7. (a) Prove that $F_n \equiv 5 \pmod{12}$ for $n \ge 1$. (b) If F_n is prime, $n \ge 1$,prove that

$$\left(\frac{3}{F_n}\right) = -1.$$

Since for F_n prime, $\left(\frac{3}{F_n}\right) \equiv 3^{(F_n-1)/2} \pmod{F_n}$, we have $3^{(F_n-1)/2} \equiv -1 \pmod{F_n}$ and we find the "only if" direction.

Conversely, suppose that $3^{(F_n-1)/2} \equiv -1 \pmod{F_n}$.Then (1) $3^{F_n-1} \equiv 1 \pmod{F_n}$, and (2) for any prime divisor p of $F_n - 1$, we have $3^{(F_n-1)/p} \not\equiv 1$. (The only p is 2, of course.) We apply the Lucas-Lehmer primality test and conclude that F_n is prime. □

[27] From http://www.prothsearch.com/fermat.html, compiled by Wilfrid Keller.

The point of this theorem is that in practice one can calculate $3^{(F_n-1)/2} \pmod{F_n}$ very rapidly. We have

$$3,\ 3^2,\ 3^{2^2} = (3^2)^2, \dots, 3^{2^{k+1}} = (3^{2^k})^2, \dots$$

Thus, we only need to do about 2^n squaring operations, *done mod* F_n, to find $3^{(F_n-1)/2} \pmod{F_n}$. On the other hand, brute force search for a divisor of $2^{2^n}+1$ would require trying all integers up to $\sqrt{2^{2^n}} = 2^{2^{n-1}}$ which is exponentially bigger than 2^n.

For example, try running PARI on `x = Mod(3, 65537);for(n=1, 15, x = x^2);print(x)` and on `x = Mod(3, 4294967297);for(n=1, 31, x = x^2);print(x)` to verify that F_4 is prime and F_5 is composite. Even for $F_7 = 2^{128}+1 = 340282366920938463463374607431768211457$ the computation is essentially instantenuous.

Besides the nice connection to ancient problems in geometry, the Fermat numbers illustrate an important point. Finding large primes is not easy, and for that numbers that are of a special form are useful. This is a problem of much relevance to cryptography.

The problem of factoring an integer is a hard problem, believed to be not in the complexity class P. In contrast, the problem of deciding whether a given integer n is a prime or not is in the complexity class P, meaning there is a polynomial time algorithm to decide it. Clearly, this algorithm only provides the answer yes/no to the question *"is n prime?"* and not the factorization of n. This result, published in 2004, created quite a sensation and is due to Agrawal, Kayal and Saxena.[28] However, there is a big difference between an algorithm that is theoretically in P and actual real-life verification. For integers of a special form, and Fermat numbers are a good example, there are tailor-made fast methods to prove, or disprove, primality.

14. THE PROJECTIVE VARIETY $a_0x_0^m + a_1x_1^m + \cdots + a_nx_n^m = 0$

Our main objective in this section is to study the number of points on a very particular non-singular projective hypersurface. We are interested in the quantity

$$N(a_0x_0^m + a_1x_1^m + \cdots + a_nx_n^m = 0),$$

where the a_i are non-zero scalars of a finite field $\mathbb{F} = \mathbb{F}_q$ and the number of solutions is calculated in the projective space $\mathbb{P}^n(\mathbb{F})$. Some times, since the equation could be also thought of as an equation in $\mathbb{A}^{n+1}(\mathbb{F})$, we will emphasize and write

$$N^{\mathrm{proj}}(a_0x_0^m + a_1x_1^m + \cdots + a_nx_n^m = 0), \quad N^{\mathrm{aff}}(a_0x_0^m + a_1x_1^m + \cdots + a_nx_n^m = 0),$$

to distinguish between the two. A basic observation is that

$$N^{\mathrm{proj}} = \frac{N^{\mathrm{aff}} - 1}{q-1}.$$

Indeed, any solution $(x_0, \dots, x_n)$ to the affine equation, except for the zero solution, defines a solution $[x_0 : \cdots : x_n]$ to the projective equation, and the map $(x_0, \dots, x_n) \mapsto [x_0 : \cdots : x_n]$ is $(q-1) : 1$.

The final answer to the problem will be a complicated formula involving Jacobi sums which generalize Gauss sums. To assist the reader a "cheat-sheet" is provided in Appendix A. To motivate and explain the main idea, we begin with some examples.

[28] M. Agrawal, N. Kayal and N. Saxena: PRIMES is in P. Ann. of Math. (2) 160 (2004), no. 2, 781–793.

14.1. **A motivating example.** Let $p > 2$ be a prime, $q = p^s$, and let $\lambda^{(s)}$ be the generalized Legendre symbol for $a \in \mathbb{F}_q$:

$$\lambda^{(s)}(a) = \begin{cases} 1, & a \neq 0 \text{ is a square in } \mathbb{F}_q; \\ -1, & a \neq 0 \text{ is not a square in } \mathbb{F}_q; \\ 0, & a = 0. \end{cases}$$

We have seen that $\lambda^{(s)}$ is a character of (exact) order 2, and, in fact,

$$\lambda^{(s)}(a) = \left(\frac{\mathrm{Nm}_{\mathbb{F}/\mathbb{F}_p}(a)}{p} \right).$$

The notation $\lambda^{(s)}$ is convenient for typographical reasons, but a more suggestive notation is

$$\left(\frac{a}{p} \right)^{(s)} := \left(\frac{\mathrm{Nm}_{\mathbb{F}/\mathbb{F}_p}(a)}{p} \right). \tag{13}$$

Note that we have

$$\left(\frac{a}{p} \right)^{(s)} = a^{\frac{q-1}{2}},$$

in the Field $\mathbb{F}_q$.

Example 14.1.1. $N(x^2 + y^2 = 1)$ over $\mathbb{F}_q$.

We perform the calculation below. The final answer gives the main term q and an error term expressed in terms of the extended Legendre symbol (namely, the character $\lambda^{(s)}$). Of course, at a later point we would want to justify using the terminology "error term" by showing that it is actually small than the "main term".

$$\begin{aligned} N(x^2 + y^2 = 1) &= \sum_{a+b=1} N(x^2 = a) \cdot N(y^2 = b) \\ &= \sum_{a+b=1} \left(1 + \left(\frac{a}{p} \right)^{(s)} \right) \left(1 + \left(\frac{b}{p} \right)^{(s)} \right) \\ &= q + \sum_{a} \left(\frac{a}{p} \right)^{(s)} + \sum_{b} \left(\frac{b}{p} \right)^{(s)} + \sum_{a+b=1} \left(\frac{a}{p} \right)^{(s)} \left(\frac{b}{p} \right)^{(s)} \\ &= q + \sum_{a+b=1} \left(\frac{a}{p} \right)^{(s)} \left(\frac{b}{p} \right)^{(s)}. \end{aligned}$$

In the calculation we used that, by Lemma 13.1.4, $\sum_a \left(\frac{a}{p} \right)^{(s)} = 0$. Note the error term:

$$\boxed{\sum_{a+b=1} \left(\frac{a}{p} \right)^{(s)} \left(\frac{b}{p} \right)^{(s)}}$$

Example 14.1.2. $N(x^3 + y^3 = 1)$ over $\mathbb{F}_q$, $q = p^s$, $3 | (q-1)$.

The interesting case here is when $3|(q-1)$, in the sense that when $3 \nmid (q-1)$ the answer is much simpler. Show the following.

Exercise 14.1.3. Show that if $q \equiv 2 \pmod 3$ then

$$N(x^3 + y^3 = 1) = q.$$

Assume then that $q \equiv 1 \pmod 3$. Then $\mathbb{X}_q \cong \mu_{q-1}$ is a cyclic group whose order is divisible by 3 and so $\mathbb{X}_q[3]$, the subgroup consisting of characters χ of $\mathbb{F}_q^\times$ such that $\chi^3 = \epsilon$, is a cyclic group of order 3. Say,

$$\mathbb{X}_q[3] = \{\epsilon, \chi, \chi^2\}, \quad \chi^3 = \epsilon.$$

Proposition 13.2.1 says that

$$N(x^3 = a) = \epsilon(a) + \chi(a) + \chi^2(a).$$

We use that to calculate $N(x^3 + y^3 = 1)$.

$$\begin{aligned} N(x^3 + y^3 = 1) &= \sum_{a+b=1} N(x^3 = a) \cdot N(y^3 = b) \\ &= \sum_{a+b=1} \left(\sum_{i=0}^{2} \chi^i(a) \right) \left(\sum_{i=0}^{2} \chi^i(b) \right) \\ &= \sum_{i,j=0}^{2} \sum_{a+b=1} \chi^i(a)\chi^j(b) \\ &= q + \sum_{\substack{i,j=0 \\ (i,j)\neq(0,0)}}^{2} \sum_{a+b=1} \chi^i(a)\chi^j(b) \end{aligned}$$

Once more, note the expressions that appear in this formula:

$$\boxed{\textstyle\sum_{a+b=1} \chi^i(a)\chi^j(b)}$$

14.2. **Jacobi sums.** Motivated by the two examples just discussed, we introduce Jacobi sums.

Let $\chi_1, \dots, \chi_\ell$, be characters of $\mathbb{F}_q^\times$, extended to $\mathbb{F}_q$. Now $q = p^s$, where p is any prime, including the prime 2. Define two **Jacobi sums** by the following formulas. (The first definition is classic, the second convenient; we follow Ireland & Rosen here).

$$J(\chi_1, \dots, \chi_\ell) = \sum_{t_1+\cdots+t_\ell=1} \chi_1(t_1) \cdots \chi_\ell(t_\ell), \tag{14}$$

$$J_0(\chi_1, \dots, \chi_\ell) = \sum_{t_1+\cdots+t_\ell=0} \chi_1(t_1) \cdots \chi_\ell(t_\ell). \tag{15}$$

In both definitions the $t_i \in \mathbb{F}_q$. Note that for any permutation $\sigma \in S_\ell$ we have

$$J(\chi_{\sigma(1)}, \dots, \chi_{\sigma(\ell)}) = J(\chi_1, \dots, \chi_\ell),$$

and similarly for J_0. Finally, we remark that the case $\ell = 1$ is allowed but then

$$J(\chi) = 1, \quad J_0(\chi) = \delta_{\chi,\epsilon} \tag{16}$$

(namely, $J_0(\chi) = 1$ if $\chi = \epsilon$, and 0 otherwise).

Revisiting the examples above, in this notation we have the formulas:

Example 14.2.1. $N(x^2+y^2=1)$ over $\mathbb{F}_q$, $q=p^s$, p odd.

$$N(x^2+y^2=1) = q + J(\lambda^{(s)},\lambda^{(s)}), \quad \lambda^{(s)}(a) = \left(\frac{a}{p}\right)^{(s)}.$$

Example 14.2.2. $N(x^3+y^3=1)$ over $\mathbb{F}_q$, $q=p^s$, $3|(q-1)$. Let χ be a generator of $\mathbb{X}_q[3]$. Then,

$$N(x^3+y^3=1) = q + \sum_{\substack{i,j=0 \\ (i,j)\neq(0,0)}}^{2} J(\chi^i,\chi^j).$$

14.2.1. *The relation between J and J_0.* The next proposition establishes some first properties of Jacobi sums and shows that the introduction of the Jacobi sum variant J_0 is mostly for notational convenience.

Proposition 14.2.3. *The Jacobi sums of ℓ characters have the following properties:*

(1) $J(\epsilon,\dots,\epsilon) = J_0(\epsilon,\dots,\epsilon) = q^{\ell-1}$.

(2) If at least one, but not all, of the χ_i equals ϵ,

$$J(\chi_1,\dots,\chi_\ell) = J_0(\chi_1,\dots,\chi_\ell) = 0.$$

(3) If $\chi_\ell \neq \epsilon$ then

$$J_0(\chi_1,\dots,\chi_\ell) = \begin{cases} 0, & \chi_1\chi_2\cdots\chi_\ell \neq \epsilon; \\ \chi_\ell(-1)(q-1)J(\chi_1,\dots,\chi_{\ell-1}), & \chi_1\chi_2\cdots\chi_\ell = \epsilon. \end{cases}$$

Example 14.2.4. Using the second property of the Proposition we find the simplified formula

$$N(x^3+y^3=1) = q + J(\chi,\chi) + 2J(\chi,\chi^2) + J(\chi^2,\chi^2).$$

Proof of Proposition 14.2.3. The case $\ell = 1$ is immediate. Assume therefore that $\ell \geq 2$.

The first part is clear; it amounts to $N(t_1+\cdots+t_\ell=\alpha) = q^{\ell-1}$, for $\alpha = 0$ or 1, and this is in fact true for any $\alpha \in \mathbb{F}_q$.

For part (2), by symmetry we may suppose that $\chi_\ell = \epsilon$ and $\chi_1 \neq \epsilon$. Then, for $\alpha = 0$ or 1, we have

$$\begin{aligned} \sum_{t_1+\cdots+t_\ell=\alpha} \chi_1(t_1)\cdots\chi_\ell(t_\ell) &= \sum_{t_1,\dots,t_{\ell-1}} \chi_1(t_1)\cdots\chi_{\ell-1}(t_{\ell-1}) \\ &= \left(\sum_{t_1}\chi_1(t_1)\right)\left(\sum_{t_2,\dots,t_{\ell-1}} \chi_2(t_2)\cdots\chi_{\ell-1}(t_{\ell-1})\right) \\ &= 0, \end{aligned}$$

by Lemma 13.1.4.

For part (3) we have the following calculation. Note that in the first equality below we may assume that $s \neq 0$ because, as $\chi_\ell \neq \epsilon$, $\chi_\ell(0) = 0$. In the second equality we use the substitution $t_i = -st_i'$ and then $\chi_j(t_i) = \chi_j(-1)\chi_j(s)\chi_j(t_i')$. In the last equality we use again Lemma 13.1.4.

$$\begin{aligned}
J_0(\chi_1,\dots,\chi_\ell) &= \sum_{s\neq 0}\left(\sum_{t_1+\dots+t_{\ell-1}=-s}\chi_1(t_1)\cdots\chi_{\ell-1}(t_{\ell-1})\right)\cdot\chi_\ell(s)\\
&=(\chi_1\cdots\chi_{\ell-1})(-1)\sum_{s\neq 0}\left(\sum_{t_1'+\dots+t_{\ell-1}'=1}(\chi_1\cdots\chi_{\ell-1})(s)\chi_1(t_1')\cdots\chi_{\ell-1}(t_{\ell-1}')\right)\cdot\chi_\ell(s)\\
&=(\chi_1\cdots\chi_{\ell-1})(-1)\left(\sum_{s\neq 0}(\chi_1\cdots\chi_\ell)(s)\right)\cdot J(\chi_1,\dots,\chi_{\ell-1})\\
&=\begin{cases}0, & \chi_1\cdots\chi_\ell\neq\epsilon;\\ \chi_\ell(-1)\cdot(q-1)\cdot J(\chi_1,\dots,\chi_{\ell-1}), & \chi_1\cdots\chi_\ell=\epsilon.\end{cases}
\end{aligned}$$

□

14.3. **Gauss and Jacobi sums.** Our main goal here is to produce a connection between Gauss and Jacobi sums. The formulas and proofs are rather mind-numbing, so a word of motivation might be required. We have seen that Jacobi sums appear in the context of counting the number of solutions N for an equation over a finite fields. However, their definition is involved and, at this point, it is not clear if the terms in N that involve Jacobi sums should be regarded as error terms, or main terms. Namely, we would like to have an estimate on the size of Jacobi sums. When we connect Jacobi sums to Gauss sums, we'd be able to use the estimates we have on Gauss sums to conclude estimates for Jacobi sums and thereby for N. This strategy goes a long way and can be used for much more general systems of equations then the simple equations we are considering. And, vice-versa, estimates coming from Weil's conjectures, allow to produce estimates to general trigonometric sums, and in particular for Jacobi sums.[29]

Theorem 14.3.1. *Assume that $\chi_1,\dots,\chi_\ell$, $\ell\geq 2$, are all non-trivial characters in $\mathbb{X}_q$, i.e. not equal to ϵ.*

(1) If $\chi_1\cdots\chi_\ell\neq\epsilon$ then

$$\mathfrak{g}(\chi_1)\cdots\mathfrak{g}(\chi_\ell)=J(\chi_1,\dots,\chi_\ell)\cdot\mathfrak{g}(\chi_1\cdots\chi_\ell).$$

(2) If $\chi_1\cdots\chi_\ell=\epsilon$, then

$$\mathfrak{g}(\chi_1)\cdots\mathfrak{g}(\chi_\ell)=q\cdot\chi_\ell(-1)\cdot J(\chi_1,\dots,\chi_{\ell-1}),$$

and

(3)

$$J(\chi_1,\dots,\chi_\ell)=-\chi_\ell(-1)\cdot J(\chi_1,\dots,\chi_{\ell-1}).$$

Proof. We begin with (1).

[29] A good introduction to these high-powered techniques is the article by Nicholas M. Katz, Sommes exponentielles *Astérisque*, 79 (1980). Unfortunately, it assumes much background from the reader. This text also adds Fourier analysis into the mix; a point of view that provides a much more conceptual understanding of Gauss sums, as Fourier transforms, and Jacobi sums as a scaling factor between the product and convolution of characters.

$$\begin{aligned}
\mathfrak{g}(\chi_1)\cdots\mathfrak{g}(\chi_\ell) &= \left(\sum_{t_1}\chi_1(t_1)\psi(t_1)\right)\cdots\left(\sum_{t_\ell}\chi_\ell(t_\ell)\psi(t_\ell)\right)\\
&= \sum_{t_1,\ldots,t_\ell}\chi_1(t_1)\cdots\chi_\ell(t_\ell)\psi(t_1+\cdots+t_\ell)\\
&= \sum_{s\in\mathbb{F}}\psi(s)\sum_{t_1+\cdots+t_\ell=s}\chi_1(t_1)\cdots\chi_\ell(t_\ell)\\
&\overset{1}{=} \sum_{s\in\mathbb{F}^\times}\psi(s)\sum_{t_1+\cdots+t_\ell=s}\chi_1(t_1)\cdots\chi_\ell(t_\ell)\\
&= \sum_{s\in\mathbb{F}^\times}\psi(s)(\chi_1\cdots\chi_\ell)(s)\sum_{t_1+\cdots+t_\ell=1}\chi_1(t_1)\cdots\chi_\ell(t_\ell)\\
&= \sum_{s\in\mathbb{F}^\times}\psi(s)(\chi_1\cdots\chi_\ell)(s)J(\chi_1,\ldots,\chi_\ell)\\
&= \mathfrak{g}(\chi_1\cdots\chi_\ell)\cdot J(\chi_1,\ldots,\chi_\ell).
\end{aligned}$$

The justification for the equality marked by 1 is that $J_0(\chi_1,\ldots,\chi_\ell)=0$ by Proposition 14.2.3.

Proceeding to (2), we will use (1) for $\chi_1\ldots\chi_{\ell-1}=\chi_\ell^{-1}=\bar{\chi}_\ell\neq\epsilon$. We have,

$$\mathfrak{g}(\chi_1)\cdots\mathfrak{g}(\chi_{\ell-1})=\mathfrak{g}(\chi_1\cdots\chi_{\ell-1})\cdot J(\chi_1,\ldots,\chi_{\ell-1})=\mathfrak{g}(\bar{\chi}_\ell)J(\chi_1,\ldots,\chi_{\ell-1}).$$

Multiply this equality by $\mathfrak{g}(\chi_\ell)$ and use that by Corollary 13.3.4 $\mathfrak{g}(\chi_\ell)\mathfrak{g}(\bar{\chi}_\ell)=\chi_\ell(-1)q$ to conclude

$$\mathfrak{g}(\chi_1)\cdots\mathfrak{g}(\chi_{\ell-1})\mathfrak{g}(\chi_\ell)=q\chi_\ell(-1)J(\chi_1,\ldots,\chi_{\ell-1}).$$

We now prove part (3). We consider the case $\ell=2$ and $\ell>2$ separately. In the case $\ell=2$, let $\chi:=\chi_1\neq\epsilon$. Then,

$$J(\chi_1,\chi_2)=J(\chi,\chi^{-1})=\sum_{a+b=1}\chi(a)\chi^{-1}(b)=\sum_{a+b=1,b\neq0}\chi(a/b).$$

Note that b is determined by a and the condition $b\neq0$ amount to $a\neq1$. Also note that the image of $\mathbb{F}-\{1\}$ under the map $a\mapsto a/(1-a)$ is $\mathbb{F}-\{-1\}$. Therefore, we find that

$$J(\chi,\chi^{-1})=\sum_{a\neq1}\chi(a/(1-a))=\sum_{c\neq-1}\chi(c)=\sum_c\chi(c)-\chi(-1)=-\chi(-1)=-\chi(-1)J(\chi).$$

For future reference we record this:

(17)
$$\boxed{J(\chi,\chi^{-1})=-\chi(-1),\quad \chi\neq\epsilon}$$

Suppose now that $\ell>2$. In proving part (1) we calculated that

$$\mathfrak{g}(\chi_1)\cdots\mathfrak{g}(\chi_\ell)=\sum_{s\in\mathbb{F}}\psi(s)\sum_{t_1+\cdots+t_\ell=s}\chi_1(t_1)\cdots\chi_\ell(t_\ell).$$

Now we find, using calculations as in part (1) and $\chi_1\cdots\chi_\ell=1$:

$$\begin{aligned}
\mathfrak{g}(\chi_1)\cdots\mathfrak{g}(\chi_\ell) &= J_0(\chi_1,\ldots,\chi_\ell)+\sum_{s\in\mathbb{F}^\times}\psi(s)\sum_{t_1+\cdots+t_\ell=s}\chi_1(t_1)\cdots\chi_\ell(t_\ell)\\
&= J_0(\chi_1,\ldots,\chi_\ell)+J(\chi_1,\ldots,\chi_\ell)\sum_{s\in\mathbb{F}^\times}\psi(s)
\end{aligned}$$

Now, using Lemma 13.1.1 and Proposition 14.2.3, it follows that

$$\begin{aligned}\mathfrak{g}(\chi_1)\cdots\mathfrak{g}(\chi_\ell) =& J_0(\chi_1,\dots,\chi_\ell) - J(\chi_1,\dots,\chi_\ell)\\ =& \chi_\ell(-1)(q-1)J(\chi_1,\dots,\chi_{\ell-1}) - J(\chi_1,\dots,\chi_\ell).\end{aligned}$$

Using part (2), it follows that

$$q\chi_\ell(-1)J(\chi_1,\dots,\chi_{\ell-1}) = \chi_\ell(-1)(q-1)J(\chi_1,\dots,\chi_{\ell-1}) - J(\chi_1,\dots,\chi_\ell),$$

and from which that

$$J(\chi_1,\dots,\chi_\ell) = -\chi_\ell(-1)J(\chi_1,\dots,\chi_{\ell-1}).$$

□

14.3.1. *Absolute value of Jacobi sums.* As a corollary of Theorem 14.3.1 we get the following information about the size of Jacobi sums.

Corollary 14.3.2. *Assume that $\chi_1,\dots,\chi_\ell$ are non-trivial characters of $\mathbb{F}_q$.*

(1) If $\chi_1\cdots\chi_\ell \neq \epsilon$ then

$$\left|J(\chi_1,\dots,\chi_\ell)\right| = q^{(\ell-1)/2}.$$

(2) If $\chi_1\cdots\chi_\ell = \epsilon$ then

$$\left|J(\chi_1,\dots,\chi_\ell)\right| = q^{(\ell-2)/2}$$

and

$$\left|J_0(\chi_1,\dots,\chi_\ell)\right| = (q-1)q^{(\ell-2)/2}.$$

Proof. The first claim follows from part (1) of Theorem 14.3.1 and the estimate $|\mathfrak{g}(\chi)| = \sqrt{q}$ for $\chi \neq \epsilon$.

The first estimate in part (2) follows from parts (2) & (3) of the same Theorem. The second estimate follows from Proposition 14.2.3 and the first claim. □

14.4. **Application of Jacobi sums to $p = a^2 + b^2$.** We use Jacobi sums to deduce a classical theorem of Fermat concerning which primes can be written as a sum of squares. Note that $2 = 1^2 + 1^2$ and a prime p congruent to 3 mod 4 cannot be a sum of two squares since any square mod 4 is either 0 or 1. Thus, the only interested case is for primes that are congruent to 1 mod 4.

Theorem 14.4.1 (P. de Fermat)**.** *Let $p \equiv 1 \pmod 4$ be a prime then p is the sum of two square integers:*

$$p = a^2 + b^2.$$

Proof. As $p \equiv 1 \pmod 4$, $\mathbb{X}_p$ is a group of order divisible mod 4 and so there is a character χ of $\mathbb{F}_p^\times$ of order 4. Such a character necessarily takes the values $\{\pm 1, \pm i\}$. Now,

$$J(\chi,\chi) = \sum_{t_1+t_2=1} \chi(t_1)\chi(t_2) = a + bi,$$

for some integers a, b. But, by Corollary 14.3.2,

$$a^2 + b^2 = \left|J(\chi,\chi)\right|^2 = p.$$

□

It will be a mistake to think that this is the technique to prove similar statements about primes appearing as quadratic expressions in integers. The right techniques come from algebraic number theory.[30] Nonetheless, a similar argument can be used to solve the following exercise.

Exercise 14.4.2. ★ Prove that if $p \equiv 1 \pmod 3$ then for suitable integers a, b

$$p = a^2 - ab + b^2.$$

14.5. **Application of Jacobi sums to $N(x_1^2 + \cdots + x_\ell^2 = 1)$.** Assume that $p > 2$. Let us use our results about Jacobi sums to understand better the number of solution to the quadraric affine hypersuface $x_1^2 + \cdots + x_\ell^2 = 1$ in $\mathbb{F}_q$. Recall the notation (13) of the extended Legendre symbol $\left(\frac{a}{p}\right)^{(s)}$ that we also denote here temporarily χ (because of typographical reasons), and recall also Example 13.2.2. We have then

$$\begin{aligned} N(x_1^2 + \cdots + x_\ell^2 = 1) &= \sum_{a_1+\cdots+a+r=1} N(x_1^2 = a_1) \cdots N(x_e ll^2 = a_\ell) \\ &= \sum_{a_1+\cdots+a+r=1} \left(1 + \left(\frac{a_1}{p}\right)^{(s)}\right) \cdots \left(1 + \left(\frac{a_\ell}{p}\right)^{(s)}\right) \\ &= \sum_{(i_1,\ldots,i_\ell)\in\{0,1\}^\ell} J(\chi^{i_1}, \ldots, \chi^{i_\ell}). \end{aligned}$$

If some $i_j = 0$, but not all, then $J(\chi^{i_1}, \ldots, \chi^{i_\ell}) = 0$ by Proposition 14.2.3. As well, $J(\epsilon, \ldots, \epsilon) = q^{\ell-1}$. Thus,

$$N(x_1^2 + \cdots + x_\ell^2 = 1) = q^{\ell-1} + J(\chi, \ldots, \chi).$$

In the Jacobi symbol, χ appears ℓ times and so we need to distinguish two cases:

- If ℓ is odd, part (1) of Theorem 14.3.1 gives $J(\chi, \ldots, \chi) = (\mathfrak{g}(\chi)^2)^{\frac{\ell-1}{2}} = \chi(-1)^{\frac{\ell-1}{2}} q^{\frac{\ell-1}{2}}$, where we have used also Corollary 13.3.4.
- If ℓ is even, then, by the same results, we have $J(\chi, \ldots, \chi) = \frac{-1}{q}(\mathfrak{g}(\chi)^2)^{\frac{\ell}{2}} = -\chi(-1)^{\ell/2} q^{\frac{\ell-2}{2}}$.

Theorem 14.5.1. *Suppose that p is an odd prime. Let $q = p^s$ and χ the extended Legendre symbol $\lambda^{(s)}$. The number of solutions to $x_1^2 + \cdots + x_\ell^2 = 1$ in $\mathbb{F}_q$ is:*

$$N(x_1^2 + \cdots + x_\ell^2 = 1) = q^{\ell-1} + \begin{cases} \chi(-1)^{\frac{\ell-1}{2}} q^{\frac{\ell-1}{2}}, & \ell \text{ odd} \\ -\chi(-1)^{\frac{\ell}{2}} q^{\frac{\ell-2}{2}}, & \ell \text{ even.} \end{cases}$$

We have $q = p^s$ and $\chi(a) = a^{\frac{q-1}{2}}$. If s is even, $\chi(-1) = 1$, while if s is odd, $\chi(-1) = (-1)^{\frac{p-1}{2}}$. Thus, in both cases $\chi(-1) = (-1)^{\frac{s(p-1)}{2}}$. That is, the formulas in the Theorem are completely explicit and depend on the parity of s, the parity of ℓ and $p \pmod 4$.

Exercise 14.5.2. Prove that for $p \equiv 1 \pmod 3$ we have

$$N(x^3 + y^3 = 1) = p - 2 + 2\mathrm{Re}(J(\chi, \chi)),$$

where we are considering solutions over $\mathbb{F}_p$ and χ is a non-trivial cubic character. Your starting point should be Example 14.2.4.

[30] An excellent book in this direction is David A, Cox, *Primes of the form $x^2 + ny^2$. Fermat, class field theory, and complex multiplication.*

In the exercise, the exact answer remains unclear since although we know the absolute value of $J(\chi,\chi)$ (which is $\sqrt{p}$), we do not know its real part.

Exercise 14.5.3. Let $A = 2\mathrm{Re}J(\chi,\chi)$. Prove that $J(\chi,\chi) = a + b\omega$ where $\omega = \frac{-1+\sqrt{-3}}{2}$ and $a, b \in \mathbb{Z}$. Conclude that $A = 2a - b$ and $a^2 - ab + b^2 = p$. Furthermore, prove that if we let $B = b/3$ then

$$4p = A^2 + 27B^2.$$

In the last exercise, in fact one can prove that B is an integer and $A \equiv 1 \pmod 3$, but this is a bit subtle. Furthermore, one can show that there is a solution in integers to

$$4p = A^2 + 27B^2,$$

satisfying $A \equiv 1 \pmod 3$ and this determines the solution uniquely up to replacing B by $-B$. Thus, modulo these missing pieces, one concludes a theorem of Gauss.

Theorem 14.5.4 (Gauss)**.** *Let $p \equiv 1 \pmod 3$ be a prime number. There are integers A, B such that $4p = A^2 + 27B^2$ and $A \equiv 1 \pmod 3$; moreover, this determines A uniquely. In terms of these,*

$$N(x^3 + y^3 = 1) = p - 2 + A.$$

Exercise 14.5.5. For $p = 7, 13$ calculate by hand the points on $x^3 + y^3 = 1$ over $\mathbb{F}_p$, as well as A and verify Gauss's Theorem.

Using Gauss theorem, find the number of solutions for the equation $x^3 + y^3 = 1$ for $p = 97$.

Exercise 14.5.6. Does it ever happen for $p \equiv 1 \pmod 3$ that $N(x^3 + y^3 = 1) = p$?, what about $N(x^3 + y^3 = 1) = p - 1$? Suppose that p and $p - 2$ are primes, can it happen that the number of solutions to $x^3 + y^3 = 1 \bmod p$ is the same as the number of solutions mod $p - 2$? Explain how to find large p for which A is close to $2\sqrt{p}$ (and thus $N(x^3 + y^3 = 1)$ is very close to the maximum possible number of points allowed by the Hasse bound $p + 2\sqrt{p}$).

14.6. **Application of Jacobi sums to $N(a_1x_1^{\ell_1} + \cdots + a_rx_r^{\ell_r} = b)$.** We now apply Jacobi sums to the more difficult problem of calculating $N(a_1x_1^{\ell_1} + \cdots + a_rx_r^{\ell_r} = b)$. The answer is complicated, but it gives a very good estimate.

In considering the equation $a_1x_1^{\ell_1} + \cdots + a_rx_r^{\ell_r} = b$, we make the following assumptions:

- $a_i, b \in \mathbb{F}_q$, $a_i \neq 0$ for all i.
- $\ell_i | (q-1)$ for all i.

Exercise 14.6.1. The second condition is made for convenience only. More precisely, let $d_i = \gcd(\ell_i, q-1)$. Prove that

$$N(a_1x_1^{\ell_1} + \cdots + a_rx_r^{\ell_r} = b) = N(a_1x_1^{d_1} + \cdots + a_rx_r^{d_r} = b).$$

(Use that if $(r, q-1) = 1$, the map $x \mapsto x^r$ is an isomorphism of groups $\mathbb{F}_q^\times \to \mathbb{F}_q^\times$.)

We begin our analysis "as usual", making use of Proposition 13.2.1.

$$\begin{aligned}N(a_1x_1^{\ell_1}+\cdots+a_rx_r^{\ell_r}=b) &= \sum_{t_1+\cdots+t_r=b} N(a_1x_1^{\ell_1}=t_1)\cdots N(a_rx_r^{\ell_r}=t_r)\\ &= \sum_{a_1t_1+\cdots+a_rt_r=b} N(x_1^{\ell_1}=t_1)\cdots N(x_r^{\ell_r}=t_r)\\ &= \sum_{a_1t_1+\cdots+a_rt_r=b}\left(\sum_{\chi_1\in\mathbb{X}_q[\ell_1]}\chi_1(t_1)\right)\cdots\left(\sum_{\chi_1\in\mathbb{X}_q[\ell_1]}\chi_1(t_1)\right)\\ &= \sum_{(\chi_1,\dots,\chi_r)\in\mathbb{X}_q[\ell_1]\times\cdots\times\mathbb{X}_q[\ell_r]}\left(\sum_{a_1t_1+\cdots+a_rt_r=b}\chi_1(t_1)\cdots\chi_r(t_r)\right).\end{aligned}$$

We next express the expressions appearing the parentheses in terms of Jacobi sums. We distinguish two cases.

- For $b=0$,

$$\begin{aligned}\sum_{a_1t_1+\cdots+a_rt_r=0}\chi_1(t_1)\cdots\chi_r(t_r) &= \chi_1(a_1^{-1})\cdots\chi_r(a_r^{-1})\sum_{t_1+\cdots+t_r=0}\chi_1(t_1)\cdots\chi_r(t_r)\\ &= \chi_1(a_1^{-1})\cdots\chi_r(a_r^{-1})J_0(\chi_1,\dots,\chi_r).\end{aligned}$$

  We recall that by Proposition 14.2.3 $J_0(\chi_1,\dots,\chi_r)$ is equal to:
  (1) q^{r-1}, if $\chi_1=\cdots=\chi_r=\epsilon$;
  (2) 0, if some $\chi_i=\epsilon$, but not all;
  (3) 0, if some $\chi_i\neq\epsilon$ and $\chi_1\cdots\chi_r\neq\epsilon$.
  Let us introduce the notation

$$\mathbb{X}_q[\ell]^*=\mathbb{X}_q[\ell]\setminus\{\epsilon\}.$$

  We conclude that

$$\begin{aligned}N(a_1x_1^{\ell_1}+\cdots+a_rx_r^{\ell_r}=0) &= q^{r-1}+\sum_{\substack{(\chi_1,\dots,\chi_r)\in\mathbb{X}_q[\ell_1]^*\times\cdots\times\mathbb{X}_q[\ell_r]^*\\ \chi_1\cdots\chi_r=\epsilon}}\chi_1(a_1^{-1})\cdots\chi_r(a_r^{-1})J_0(\chi_1,\dots,\chi_r).\end{aligned}\tag{18}$$

- For $b\neq 0$,

$$\begin{aligned}\sum_{a_1t_1+\cdots+a_rt_r=b}\chi_1(t_1)\cdots\chi_r(t_r) &= (\chi_1\cdots\chi_r)(b)\chi_1(a_1^{-1})\cdots\chi_r(a_r^{-1})\sum_{t_1+\cdots+t_r=1}\chi_1(t_1)\cdots\chi_r(t_r)\\ &= (\chi_1\cdots\chi_r)(b)\chi_1(a_1^{-1})\cdots\chi_r(a_r^{-1})J(\chi_1,\dots,\chi_r).\end{aligned}$$

  Similar consideration give us that for $b\neq 0$,

$$\begin{aligned}N(a_1x_1^{\ell_1}+\cdots+a_rx_r^{\ell_r}=b) &= q^{r-1}+\sum_{(\chi_1,\dots,\chi_r)\in\mathbb{X}_q[\ell_1]^*\times\cdots\times\mathbb{X}_q[\ell_r]^*}(\chi_1\cdots\chi_r)(b)\chi_1(a_1^{-1})\cdots\chi_r(a_r^{-1})J(\chi_1,\dots,\chi_r).\end{aligned}\tag{19}$$

Combining Equations (19) and (18) with Corollary 14.3.2 we deduce the following estimates; one can clearly improve them, by estimating better the number of characters involved in the sums in the equations, and also whether the characters multiply to ϵ, or not.

Proposition 14.6.2. *We have the following estimates on the number of solutions for the equation* $a_1x_1^{\ell_1} + \cdots + a_rx_r^{\ell_r} = b$ *over* $\mathbb{F}_q$.

$$\left|N(a_1x_1^{\ell_1} + \cdots + a_rx_r^{\ell_r} = b) - q^{r-1}\right| \leq \begin{cases} \ell_1 \cdots \ell_r \cdot (q-1)q^{\frac{r-2}{2}}, & b = 0; \\ \ell_1 \cdots \ell_r \cdot q^{\frac{r-1}{2}}, & b \neq 0. \end{cases} \tag{20}$$

14.7. **Application of Jacobi sums to** $N(a_0x_0^m + a_1x_1^m + \cdots + a_nx_n^m = 0)$ **in projective space.** A consequence of the results proven above is the following theorem.

Theorem 14.7.1. *Let* $m|(q-1)$, $a_i \in \mathbb{F}^\times = \mathbb{F}_q^\times$. *Consider the number of solutions to the equation*

$$a_0x_0^m + \cdots + a_nx_n^m = 0$$

in $\mathbb{P}^n(\mathbb{F})$*; we denote it* $N^{proj}(a_0x_0^m + \cdots + a_nx_n^m = 0)$. *Then,*

$$\begin{aligned} N^{proj}(a_0x_0^m + \cdots + a_nx_n^m = 0) &= q^{n-1} + q^{n-2} + \cdots + 1 \\ &\quad + \frac{1}{q-1} \sum_{\substack{\chi_i \in \mathbb{X}_q[m]^* \\ \chi_0\chi_1\cdots\chi_n = \epsilon}} \chi_0(a_0^{-1})\chi_1(a_1^{-1}) \cdots \chi_n(a_n^{-1}) J_0(\chi_0, \chi_1, \ldots, \chi_n) \\ &= q^{n-1} + q^{n-2} + \cdots + 1 \\ &\quad + \frac{1}{q} \sum_{\substack{\chi_i \in \mathbb{X}_q[m]^* \\ \chi_0\chi_1\cdots\chi_n = \epsilon}} \chi_0(a_0^{-1})\chi_1(a_1^{-1}) \cdots \chi_n(a_n^{-1}) \mathfrak{g}(\chi_0)\mathfrak{g}(\chi_1) \cdots \mathfrak{g}(\chi_n) \end{aligned}$$

Proof. To relate the Theorem to previous results, note that any solution $(y_0, \ldots, y_n)$ in $\mathbb{A}^{n+1}(\mathbb{F})$, except the zero solution, is a projective solution $[y_0 : \cdots : y_n]$. And, furthemore, any such projective solution arises precisely from $(q-1)$ affine solutions; to wit, $(\lambda y_0, \ldots, \lambda y_n)$, $\lambda \in \mathbb{F}_q^\times$. That is, the number of solutions N^{proj} to $a_0x_0^m + \cdots + a_nx_n^m = 0$ in the projective n-space $\mathbb{P}^n$ is gotten from the number of solutions N^{aff} in affine $n+1$- space $\mathbb{A}^{n+1}$ by the formula $\frac{(N^{\mathrm{aff}}(a_0x_0^m + \cdots + a_nx_n^m = 0) - 1)}{(q-1)}$. Applying (18), this number is

$$\frac{q^n - 1}{q-1} + \frac{1}{q-1} \sum_{\substack{(\chi_0, \ldots, \chi_n) \in \mathbb{X}_q[m]^* \times \cdots \times \mathbb{X}_q[m]^* \\ \chi_0 \cdots \chi_n = \epsilon}} \chi_0(a_0^{-1}) \cdots \chi_n(a_n^{-1}) J_0(\chi_0, \ldots, \chi_n).$$

The second expression is obtained from Proposition 14.2.3, combined with Theorem 14.3.1. □

Exercise 14.7.2. ★ Prove that

$$\left|N^{\mathrm{proj}}(a_0y_0^m + \cdots + a_ny_n^m = 0) - \sharp\mathbb{P}^{n-1}(\mathbb{F}_q)\right| \leq f(m) \cdot q^{\frac{n-1}{2}}.$$

where

$$f(m) = \frac{1}{m}((m-1)^{n+1} + (-1)^{n+1}(m-1)).$$

15. Rationality of certain zeta functions

Our goal in this section is to prove the rationality of the zeta function of hypersurfaces of the form

$$a_0x_0^m + \cdots + a_nx_n^m = 0, \qquad a_i \in \mathbb{F}^\times, m|(q-1).$$

As we have seen before, the condition that $m|(q-1)$ is not serious and clearly also the case where some of the $a_i = 0$ can be dealt with easily by reducing to a problem with less variables.

Given Theorem 14.7.1, which applies equally to $\mathbb{F}$ or to a finite extension $\mathbb{F}_{[s]}$ of $\mathbb{F}$ (because we can view the a_i as lying in $\mathbb{F}_{[s]}$), what remains is to understand how the formulas change under passage from the field $\mathbb{F}$ to an extension $\mathbb{F}_{[s]}$. The field $\mathbb{F}_{[s]}$ has q^s elements and the character groups $\mathbb{X}_{q^s}$ and $\mathbb{X}_q$ are not the same and neither are the Gauss and Jacobi sums. But, this is really the extent of the problem. The part of relating the different Gauss sums is rather involved and we will not do it here, referring the reader to Ireland & Rosen, Chapter 11, §4; After doing the rest of the analysis, we will be able to prove the following theorem.

Theorem 15.0.1. *Let $\mathbb{F} = \mathbb{F}_q$, $m|(q-1)$ and $a_0, \ldots, a_n \in \mathbb{F}^\times$. Let V be the projective variety in $\mathbb{P}^n$, $n \geq 1$, defined by the equation*

$$a_0x_0^m + \cdots + a_nx_n^m = 0.$$

Then,

$$\zeta_V(T) = \frac{P(T)^{(-1)^n}}{(1-T)(1-qT)\cdots(1-q^{n-1}T)}, \tag{21}$$

where $P(T)$ is the polynomial

$$\prod_{\substack{\chi_i \in \mathbb{X}_q[m]^* \\ \chi_0\chi_1\cdots\chi_n = \epsilon}} \left(1 - \frac{(-1)^{n+1}}{q}\chi_0(a_0)^{-1}\chi_1(a_1)^{-1}\cdots\chi_n(a_n)^{-1}\cdot \mathfrak{g}(\chi_0)\mathfrak{g}(\chi_1)\cdots\mathfrak{g}(\chi_n)\cdot T\right). \tag{22}$$

15.1. **A criterion for rationality.** To prove rationality of zeta functions, we first note a simple criterion for a power series of the form arising in the definition of ζ_V to be rational.

Lemma 15.1.1. *A power series*

$$\zeta(T) = \exp\Big(\sum_{s=1}^{\infty} \frac{N_s}{s}T^s\Big), \qquad N_s \in \mathbb{C}$$

is rational, in the sense that it is of the form

$$\zeta(T) = \prod_i (1-\alpha_iT) \Big/ \prod_j (1-\beta_jT)$$

for some $\alpha_i, \beta_j \in \mathbb{C}$, if and only if, for all s,

$$N_s = \sum_j \beta_j^s - \sum_i \alpha_i^s.$$

Proof. This is a computation we have essentially done before. To begin with, assume that $\zeta(T) = \prod_i(1-\alpha_iT)/\prod_j(1-\beta_jT)$. Then, on the one hand,

$$d\log\zeta = \sum_{s=1}^{\infty} N_sT^{s-1},$$

and, on the other hand,

$$d\log\zeta = d\log\left(\prod_i(1-\alpha_i T)\right) - d\log\left(\prod_j(1-\beta_j T)\right) = \sum_i d\log(1-\alpha_i T) - \sum_j d\log(1-\beta_j T)$$
$$= \sum_j\sum_{s=1}^{\infty}\beta_j^s T^{s-1} - \sum_i\sum_{s=1}^{\infty}\alpha_i^s T^{s-1} = \sum_{s=1}^{\infty}(\sum_j\beta_j^s)T^{s-1} - \sum_{s=1}^{\infty}(\sum_{s=i}\alpha_i^s)T^{s-1},$$

hence the formula for N_s.

Now, note that two functions f, g with $d\log f = d\log g$ satisfy $d\log(f/g) = 0$ and this implies that f and g differ by a scalar. If f and g are both functions whose Taylor expansions start with 1, such as is the case with the zeta function and with the function $\prod_i(1-\alpha_i T)/\prod_j(1-\beta_j T)$, then $f = g$. Thus, the argument can be reversed! □

15.2. **Relating $\mathbb{X}_q$ and $\mathbb{X}_{q^s}$.** Let $\mathbb{F}_{[s]}$ be a degree s extension of $\mathbb{F}$. Since we may view a_i as lying in $\mathbb{F}_{[s]}$, we may apply Theorem 14.7.1 to find the number of projective solutions N_s to the equation $a_0x_0^m + \cdots + a_nx_n^m = 0$ in $\mathbb{F}_{[s]}$. The answer is

$$N_s = N_s^{\mathrm{proj}}(a_0x_0^m + \cdots + a_nx_n^m = 0) \tag{23}$$
$$= q^{s(n-1)} + q^{s(n-2)} + \cdots + 1 + \frac{1}{q^s}\sum_{\substack{\chi_i\in\mathbb{X}_{q^s}[m]^*\\ \chi_0\chi_1\cdots\chi_n=\epsilon}}\chi_0(a_0^{-1})\chi_1(a_1^{-1})\cdots\chi_n(a_n^{-1})\mathfrak{g}(\chi_0)\mathfrak{g}(\chi_1)\cdots\mathfrak{g}(\chi_n).$$

Note that the characters χ are now characters of $\mathbb{F}_{[s]}$, a field with q^s elements, so they are elements of $\mathbb{X}_{q^s}$, a group that changes with s. Thus, our first task is to relate $\mathbb{X}_q$ and $\mathbb{X}_{q^s}$.

Given a character $\chi \in \mathbb{X}_q$, consider the function $\chi^{(s)}$ defined as[31]

$$\chi^{(s)}\colon \mathbb{F}_{[s]}^\times \to \mathbb{C}^\times, \quad \chi^{(s)}(a) = \chi\circ\mathrm{Nm}_{\mathbb{F}_{[s]}/\mathbb{F}}(a).$$

Lemma 15.2.1. *The map $\chi \mapsto \chi^{(s)}$ is an injective group homomorphism*

$$\mathbb{X}_q \hookrightarrow \mathbb{X}_{q^s}.$$

Under this injection, $\mathbb{X}_q[m]$ is identified with $\mathbb{X}_{q^s}[m]$.

Proof. First, $\chi^{(s)}$ is a character, i.e., group homomorphism, because it is the composition of group homomorphisms. We also need to verify that $(\chi_1\chi_2)^{(s)} = \chi_1^{(s)}\chi_2^{(s)}$ and, indeed,

$$(\chi_1\chi_2)^{(s)}(x) = (\chi_1\chi_2)(\mathrm{Nm}_{\mathbb{F}_{[s]}/\mathbb{F}}(x)) = \chi_1(\mathrm{Nm}_{\mathbb{F}_{[s]}/\mathbb{F}}(x))\chi_2(\mathrm{Nm}_{\mathbb{F}_{[s]}/\mathbb{F}}(x)) = (\chi_1^{(s)}\chi_2^{(s)})(x).$$

The map $\chi \mapsto \chi^{(s)}$ is injective because $\mathrm{Nm}_{\mathbb{F}_{[s]}/\mathbb{F}}$ is a surjective homomorphism $\mathbb{F}_{[s]}^\times \to \mathbb{F}^\times$. Since both $\mathbb{X}_q[m]$ and $\mathbb{X}_{q^s}[m]$ have m elements, we conclude the last statement of the lemma. □

Note that for $x \in \mathbb{F}$ we have

$$\chi^{(s)}(x) = \chi(\mathrm{Nm}_{\mathbb{F}_{[s]}/\mathbb{F}}(x)) = \chi(x^s) = \chi(x)^s.$$

[31] This explains our notation for the extended Legendre symbol. We denote it $\lambda^{(s)}$, or $\left(\frac{\cdot}{p}\right)^{(s)}$.

Thus, we may write,

$$N_s = N_s^{\text{proj}}(a_0x_0^m + \cdots + a_nx_n^m = 0) \tag{24}$$
$$= q^{s(n-1)} + q^{s(n-2)} + \cdots + 1 + \frac{1}{q^s} \sum_{\substack{\chi_i \in \mathbb{X}_q[m]^* \\ \chi_0\chi_1\cdots\chi_n = \epsilon}} \chi_0^{(s)}(a_0^{-1})\chi_1^{(s)}(a_1^{-1}) \cdots \chi_n^{(s)}(a_n^{-1})\mathfrak{g}(\chi_0^{(s)})\mathfrak{g}(\chi_1^{(s)}) \cdots \mathfrak{g}(\chi_n^{(s)})$$
$$= q^{s(n-1)} + q^{s(n-2)} + \cdots + 1 + \frac{1}{q^s} \sum_{\substack{\chi_i \in \mathbb{X}_q[m]^* \\ \chi_0\chi_1\cdots\chi_n = \epsilon}} \chi_0(a_0^{-1})^s\chi_1(a_1^{-1})^s \cdots \chi_n(a_n^{-1})^s\mathfrak{g}(\chi_0^{(s)})\mathfrak{g}(\chi_1^{(s)}) \cdots \mathfrak{g}(\chi_n^{(s)}).$$

15.3. **The Hasse-Davenport relation and the rationality of ζ_V.** The next task is to relate the Gauss sums $\mathfrak{g}(\chi^{(s)})$ and $\mathfrak{g}(\chi)$. The result in known as the Hasse-Davenport relation and we refer for the proof to Ireland & Rosen, Chapter 11, §4.

Theorem 15.3.1 (Hasse-Davenport)**.** *We have*

$$\mathfrak{g}(\chi^{(s)}) = -(-1)^s\mathfrak{g}(\chi)^s.$$

Substituting in Equation (24), we find that

$$N_s = N_s^{\text{proj}}(a_0x_0^m + \cdots + a_nx_n^m = 0) \tag{25}$$
$$= q^{s(n-1)} + q^{s(n-2)} + \cdots + 1 + (-1)^{n+1} \sum_{\substack{\chi_i \in \mathbb{X}_q[m]^* \\ \chi_0\chi_1\cdots\chi_n = \epsilon}} ((-1)^{n+1})^s(\frac{1}{q})^s\chi_0(a_0^{-1})^s \cdots \chi_n(a_n^{-1})^s\mathfrak{g}(\chi_0)^s \cdots \mathfrak{g}(\chi_n)^s.$$

We have written N_s exactly in the form suitable to apply the Rationality Criterion (Lemma 15.1.1). Theorem 15.0.1 has thus been proven:

$$\zeta_V(T) = \frac{P(T)^{(-1)^n}}{(1-T)(1-qT)\cdots(1-q^{n-1}T)},$$

where $P(T)$ is the polynomial

$$\prod_{\substack{\chi_i \in \mathbb{X}_q[m]^* \\ \chi_0\chi_1\cdots\chi_n = \epsilon}} \left(1 - \frac{(-1)^{n+1}}{q}\chi_0(a_0)^{-1}\chi_1(a_1)^{-1} \cdots \chi_n(a_n)^{-1} \cdot \mathfrak{g}(\chi_0)\mathfrak{g}(\chi_1) \cdots \mathfrak{g}(\chi_n) \cdot T\right).$$

Exercise 15.3.2. ★ Prove that ζ_V provided in Theorem 15.0.1 satisfies the Weil conjectures.

15.4. **Some additional examples.** The results we have developed so far already provide rather nice answers for the two exercises given further below. However, to have a really definite formula the following additional information is needed.

Assume p is an odd prime and recall that Legendre symbol on $\mathbb{F}_p$:

$$\lambda(a) = \left(\frac{a}{p}\right), \quad a \in \mathbb{F}_p.$$

This is a character of order 2 and so,

$$\mathfrak{g}(\lambda)^2 = \mathfrak{g}(\lambda)\mathfrak{g}(\bar{\lambda}) = \chi(-1) \cdot p = \left(\frac{-1}{p}\right)p.$$

Thus, $\mathfrak{g}(\lambda)$ is equal to $\pm\sqrt{p}$ if $p \equiv 1 \pmod 4$ and is equal to $\pm i\sqrt{p}$ if $p \equiv 1 \pmod 4$. The following theorem of Gauss settles this point.

Theorem 15.4.1 (Gauss)**.**

$$\mathfrak{g}(\lambda) = \begin{cases} \sqrt{p}, & p \equiv 1 \pmod 4; \\ i\sqrt{p}, & p \equiv 3 \pmod 4. \end{cases}$$

Exercise 15.4.2. Give an explicit formula for $N^{\mathrm{proj}}(x_0^2 + \cdots + x_n^2 = 0)$.

Exercise 15.4.3. Give an explicit formula for $N^{\mathrm{proj}}(a_0x_0^2 + \cdots + a_nx_n^2 = 0)$.

In contrast, it is very hard to determine exactly cubic Gauss sums, let alone Gauss sums for general characters – even in the case when $3|(p-1)$ where we are dealing with a Gauss sum of a character of $\mathbb{F}_p$. Let χ be a character of exact order 3 of $\mathbb{F}_p^\times$, then, by Theorem 14.3.1,

$$J(\chi,\chi) = \mathfrak{g}(\chi)^2/\mathfrak{g}(\chi^2).$$

But $\chi^2 = \chi^{-1} = \bar{\chi}$ and so, applying Corollary 13.3.4 and noting that $\chi(-1) = 1$, we find that

$$J(\chi,\chi) = \frac{\mathfrak{g}(\chi)^3}{p}.$$

We have seen that $J(\chi,\chi)$ can be found by finding integer solutions to the equation

$$4p = A^2 + 27B^2,$$

such that $A \equiv 1 \pmod 3$ (that arose in the study of the number of solutions to $x^3 + y^3 = 1$; see Theorem 14.5.4). However, finding the exact expression to $\mathfrak{g}(\chi)$ from this is not easy, although known.[32]

Exercise 15.4.4. Still using the notation λ for the Legendre character. Let α be any non-trivial character of $\mathbb{F}_p^\times$. Prove that

$$J(\lambda,\alpha) = \sum_{t\in\mathbb{F}} \alpha(1-t^2).$$

(Hint: use $N(x^2 = a) = 1 + \lambda(a)$.)

Exercise 15.4.5. Consider the equation $y^2 = x^3 + a$, where $a \in \mathbb{F}_p^\times$ is fixed and $p > 3$. Find an expression for $N(y^2 = x^3 + a)$. This expression will involve $J(\lambda,\alpha)$ where α is a cubic character. How does this compare with the expression for the zeta function of the projectivized curve $y^2z = x^3 + az^3$?

Exercise 15.4.6. Let $p > 2$ be a prime and consider an equation of the form

$$C\colon \quad y^2 = f(x),$$

where f is a separable polynomial in $\mathbb{F}_p[x]$ of the degree $2g+1$.

- Prove that this is a non-singular curve in $\mathbb{A}^2$.
- Check that the corresponding projective curve in $\mathbb{P}^2$, obtained by homogenizing $y^2 - f(x)$ is singular if $g > 1$. However, one can show that there is a projective non-singular curve $\tilde{C}$ (living in some higher dimensional projective space) that contains C and such that $\tilde{C} \setminus C$ consists of a single point which is moreover defined over $\mathbb{F}_p$. The genus of $\tilde{C}$ is g and that implies that

$$\zeta_{\tilde{C}}(T) = \frac{P_1(T)}{(1-T)(1-pT)},$$

[32] A convenient modern reference is D. Schipani and M. Elia, Gauss sums of cubic character over $\mathbb{F}_{p^r}$, p odd. Bull. Pol. Acad. Sci. Math. 60 (2012), no. 1, 1–19.

where $P_1 \in \mathbb{Z}[T]$ is a polynomial of degree $2g$ and constant coefficient 1.

Assuming all that show that

$$\sharp C(\mathbb{F}_p) = p + \sum_{t \in \mathbb{F}_p} \left(\frac{f(t)}{p} \right),$$

and deduce the estimate due to Burgess

$$\Big| \sum_{t \in \mathbb{F}_p} \left(\frac{f(t)}{p} \right) \Big| \leq 2g\sqrt{p}.$$

Exercise 15.4.7. Let $a, b, c \in \mathbb{F}_p$, $p > 2$,$a \neq 0, b^2 - 4ac \neq 0$. Determine the zeta function of the affine equation:

$$ax^2 + bxy + cy^2 = 0.$$

16. Cohomology and the Weil conjectures

Very often a non-singular projective variety V over a finite field arises as the reduction of a non-singular projective variety in characteristic zero. Indeed, given a set of polynomials $f_i(x_0, \ldots, x_n) \in \mathbb{F}_q[x]$, one can find a finite extension $\mathbb{L}$ of $\mathbb{Q}$, $\mathbb{L} \subset \mathbb{C}$, and polynomials $F_i(x_0, \ldots, x_n)$ whose coefficients are algebraic integers of $\mathbb{L}$, namely, they lie in $\mathcal{O}_L$ (the ring of algebraic integers of L) and a prime ideal $\mathfrak{p}$ of $\mathcal{O}_L$ such that $\mathcal{O}_L/\mathfrak{p} \cong \mathbb{F}_q$ and the polynomials F_i reduce to f_i. Denote by $\mathbb{V}$ the variety defined by the F_i and assume it is irreducible and non-singular. Let d denote its dimension. A good case to keep in mind is when $\mathbb{F}_q = \mathbb{Z}/p\mathbb{Z}$; in this case, we simply lift the coefficients of f_i, which are mod p congruence classes, to integers. Quite surprisingly this method doesn't always work. Some of the first examples were given by Serre[33]. There are also examples of non-singular projective surfaces in $\mathbb{P}^5$ that cannot be lifted to characteristic 0. But let us assume that our situation is favourable and a lift exists.

The complex points $\mathbb{V}(\mathbb{C})$ of $\mathbb{V}$ are a compact topological manifold. As such, one can associate to $\mathbb{V}(\mathbb{C})$ cohomology spaces $\{H^i(\mathbb{V}, \mathbb{Q})\}_{i=0}^{2d}$. In fact, such cohomology spaces are associated to any topological manifold. These are finite dimensional rational vector spaces constructed by topological means. They satisfy $H^0(\mathbb{V}, \mathbb{Q}) = H^{2d}(\mathbb{V}, \mathbb{Q}) = \mathbb{Q}$. They have various properties that make their computation almost axiomatic. And, they are functorial: a continuous map of manifolds $f\colon \mathbb{V} \to \mathbb{W}$ induces linear maps $f_i^*\colon H^i(\mathbb{W}, \mathbb{Q}) \to H^i(\mathbb{V}, \mathbb{Q})$, for all i. In particular, each such linear map has a trace $\mathrm{Tr}(f_i^*)$.

Suppose that $f\colon \mathbb{V} \to \mathbb{V}$ is a map that has finitely many fixed points. The Lefschetz trace formula gives

$$\#\text{ fixed points of } f = \sum_{i=0}^{d} (-1)^i \mathrm{Tr}(f_i^*). \tag{26}$$

Consider for example, an elliptic curve $E\colon y^2 = x^3 + ax + b$ and its projective model $zy^2 = x^3 + az^2x + bz^3$. Consider the map $P \mapsto f(P) := -P$, or in coordinates $(x, y) \mapsto (x, -y)$. This map has precisely four fixed points $\{(t, 0) : t^3 + at + b = 0\} \cup \{[0 : 1 : 0]\}$ and they comprise $E[2]$. We also have $H^0(E, \mathbb{Q}) = H^2(E, \mathbb{Q}) = \mathbb{Q}$ and $H^1(E, \mathbb{Q}) \cong \mathbb{Q}^2$. One can prove that f acts

[33] J.-P. Serre, Exemples de variétés projectives en caractéristique p non relevables en caractéristique zéro. Proc. Nat. Acad. Sci. U.S.A. 47 (1961), 108–109. but see R. Vakil's article for extensive history and ultimate bad news: R. Vakil, Murphy's law in algebraic geometry: badly-behaved deformation spaces. Invent. Math. 164 (2006), no. 3, 569–590.

as multiplication by -1 on $H^1(E,\mathbb{Q})$ (and so its trace is -2) and as the identity on H^0 and H^2. Thus, the alternating sum of traces is $1-(-2)+1=4$, in agreement with Lefschetz's formula.

Weil's idea was that perhaps such a cohomology theory exists also for varieties over finite fields. Initially, such cohomology theories were called Weil cohomology theories, but today they are known by the names of the techniques used for their constructions; for example, étale cohomology, crystalline cohomology, etc. The existence of such theories was proven later by A. Grothendieck and his school, with assistance from J.-P. Serre and P. Deligne.

How is all this connected to Weil's conjectures? Let $\mathbb{F}_q=\mathbb{F}_{p^s}$ be a finite field and let $\varphi(x)=x^p$ be the Frobenius map. Weil observed that much in the same way that $\mathbb{F}_q$ is the fixed points of φ^s on $\overline{\mathbb{F}}_p$, the $\mathbb{F}_q=\mathbb{F}_{p^s}$ points of V, could be thought as the $\overline{\mathbb{F}}_p$ points of V that are fixed points for φ^s, where φ is now the Frobenius morphism on $\mathbb{P}^n$ given by

$$\varphi(x_0:\cdots:x_n)=(x_0^p:\cdots:x_n^p).$$

Let $\psi=\varphi^s$. In a good cohomology theory, one should then have the formula

$$\sharp\mathbb{V}(\mathbb{F}_q)=\sharp \text{ fixed points of } \psi=\sum_{i=0}^{2d}(-1)^i\mathrm{Tr}(\psi_i^*), \tag{27}$$

where ψ_i^* are the maps induced by ψ on $H^i(V)$[34]. Many properties then follow: for example, the functional equation property is a consequence of Poincaré duality for cohomology. Likewise, the rationality of the zeta function follows readily, with

$$\deg(P_i)=\dim(H^i(V))=\dim(H^i(\mathbb{V})).$$

The "Riemann hypothesis", though, is still a very hard fact.

For example, for the projective space itself we have that $H^{2i}(\mathbb{P}^n)$ are 1-dimensional for $i=0,1,\ldots,n$ and all other cohomology groups vanish. This leads to $\zeta_{\mathbb{P}^n}=\frac{1}{(1-T)(1-pT)\cdots(1-p^nT)}$. For an elliptic curve (which is a curve of genus 1), and more generally for a curve C of genus g, we have $H^0(C)$ and $H^2(C)$ are 1-dimensional, $H^1(C)$ is $2g$-dimensional and all other cohomology spaces vanish. A lot is known also about the cohomology of hypersurfaces and that is, on some conceptual level, in agreement with the fact that for very special hypersurfaces we were able to find the zeta functions.

17. In conclusion

Weil's conjectures have influenced greatly the development of number theory since their formulation in the middle of the 20-th century. Weil's idea, that such properties of zeta functions of varieties over finite fields, will follow formally from the construction of good cohomology theory for varieties over finite fields, together with reasonable conjectures as to the action of a power of the Frobenius homomorphism on such a cohomology theory, impacted the development of whole new theories in algebraic geometry, most notably étale cohomology and deformation theory. It continues to inspire much research: for a start, use of other cohomology theories is still being investigated[35]. Next, the topic of variation of zeta functions for families of varieties that vary in terms of some continuous parameters – for example, an elliptic curve

[34]We are being deliberately vague about the coefficients as to avoid the introduction of yet another mystery. One needs to choose a prime $\ell\neq p$; the $H^i(V)$ are constructed as $\mathbb{Q}_\ell$-vector spaces, where $\mathbb{Q}_\ell$ is the field of ℓ-adic numbers.

[35]See for example, Kiran S. Kedlaya, "Fourier transforms and p-adic "Weil II"", Compositio Mathematica 142 (2006), 1426-1450, and "Counting points on hyperelliptic curves using Monsky-Washnitzer cohomology", Journal of the Ramanujan Mathematical Society 16 (2001), 323-338. See also, Alan Lauder, "Deformation theory and the computation of zeta functions", Proceedings of the London Mathematical Society, Vol. 88 Part 3, (2004), 565-602.

$y^2 = x^3 + a(t)x + b(t)$, where $a(t), b(t)$ are functions of t and t varies over some parameter space – is an active area of research, as is the understanding the variation of a zeta function of the reduction mod p of variety defined over the integers, say. Here, again, a good example to keep in mind is the reduction of an elliptic curve $y^2 = x^3 + ax + b$, where $a, b \in \mathbb{Z}$, modulo various primes, where the most impressive result we have is the Sato-Tate conjecture. Analogues of such questions for general varieties are unknown and are again related to recent and current research that tries to prove that the statistics of the variation of the zeta functions, a statistics that generalizes the Sato-Tate distribution, are controlled by the theory of random matrices for a certain algebraic group determined by the family and the parameter space.

The cases we have dealt with in this part of the notes are very special and we could have dealt with them using ancient knowledge – Gauss and Jacobi sums. Yet, the final result giving the zeta function as a rational function where the polynomials have roots defined in terms of such sums is rather intricate. And, to date, our understanding of Gauss and Jacobi sums is not complete; this topic too remains a topic of current research.

LATTICES, GEOMETRY OF NUMBERS AND CODES.

18. INTRODUCTION

In this part of the course we discuss lattices in Euclidean space, linear binary codes and a connection between the two. We shall see various applications of lattices to number theory and to sphere packing. The proofs of some results, results that have the advantage of explaining the bigger picture, are unfortunately outside the scope of an undergraduate text, but their importance and elegance are such that it is worthwhile to include them even if no proof is offered.

The best advice to appreciate this chapter is to realize that we really don't understand higher-dimensional spaces. Questions like the best way to pack n-dimensional oranges, that are rather clear for the 2-dimensional case, seem clear but extremely hard for the 3-dimensional case, are largely beyond reach at the moment. Such questions cannot be approached from a naïve point of view; sophisticated tools are needed.

For lattices, a canonical reference is the resource book:

J. H. Conway and N. J. A. Sloane, *Sphere packings, lattices and groups.* Third edition. Grundlehren der Mathematischen Wissenschaften, 290. Springer-Verlag, New York, 1999.

It is not self-contained in the sense that many, perhaps most, of the results are not proven, but it has an extensive bibliography. The various applications of lattices were collected from a variety of resources, from books to research articles. Likewise, the material concerning binary codes was collected from many sources, so it is hard to suggest one particular reference. That said, some references we have often consulted are

J.-P. Serre, *A course in Arithmetic*, Graduate Texts in Mathematics, 7, Springer verlag.

N. J. N. Sloane, weight enumerators of codes, in *Combinatorics, Proceedings of the NATO Advanced Study Institute, 1974*.

N. D. Elkies, Lattices, Linear Codes, and Invariants, Part I & II, Notices AMS 47 (2000), no. 11 & 12.

19. BILINEAR FORMS, QUADRATIC FORMS AND EUCLIDEAN LATTICES

19.1. **Bilinear forms and quadratic forms.** A function

$$\langle x, y\rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R},$$

is called a **symmetric bilinear form** if it satisfies the following properties for all $x, x', y, y' \in \mathbb{R}^n, \alpha \in \mathbb{R}$:

(1) $\langle x, y\rangle = \langle y, x\rangle$.
(2) $\langle x + x', y\rangle = \langle x, y\rangle + \langle x', y\rangle$ and $\langle x, y + y'\rangle = \langle x, y\rangle + \langle x, y'\rangle$.
(3) $\langle \alpha x, y\rangle = \alpha\langle x, y\rangle = \langle x, \alpha y\rangle$.

By choosing a basis $\mathbf{u} = \{u_i : i = 1, \dots, n\}$ for $\mathbb{R}^n$ we can represent $\langle x, y\rangle$ by a symmetric real matrix

$$B = (b_{ij}) = (\langle u_i, u_j\rangle)_{i,j}$$

that has the property

$$\langle x, y\rangle = {}^t[x]_{\mathbf{u}} B [x]_{\mathbf{u}},$$

where we consider vectors in $\mathbb{R}^n$ as column vectors, and where $[x]_{\mathbf{u}} = {}^t(x_1, \ldots, x_n)$ is the vector of coordinates of x relative to the basis $\mathbf{u}$: $x = x_1u_1 + \cdots + x_nu_n$.

The matrix B is not uniquely determined. If we change the basis $\mathbf{u}$ to a basis $\mathbf{v}$, and let M be the change of basis matrix so that $[x]_{\mathbf{u}} = M[x]_{\mathbf{v}}$, then

$$\langle x, y\rangle = {}^t[x]_{\mathbf{u}} B [x]_{\mathbf{u}} = {}^t[x]_{\mathbf{v}} {}^t M B M [x]_{\mathbf{v}},$$

and so the matrix relative to the new basis is ${}^tMBM$. We say that B and ${}^tMBM$ are **similar** bilinear forms over $\mathbb{R}$.

To a symmetric bilinear form $\langle x, y\rangle$ we can associate the function

$$q\colon \mathbb{R}^n \to \mathbb{R}, \quad q(x) = \langle x, x\rangle,$$

which is a **quadratic form**. Namely, it satisfies

(1) $q(\alpha x) = \alpha^2 q(x)$.
(2) The function $(x, y) \mapsto \frac{1}{2}(q(x + y) - q(x) - q(y))$ is bilinear.

Indeed, starting from $\langle x, y\rangle$, the function in (2) is just $\langle x, y\rangle$ again. But, conversely, given a quadratic form q, if we let

$$\langle x, y\rangle = \frac{1}{2}(q(x + y) - q(x) - q(y))$$

we get a symmetric bilinear form such that $q(x) = \langle x, x\rangle$. As before, choosing a basis $\mathbf{u}$, and defining a matrix $B = (b_{ij})$ as above, allows us to express q explicitly. If $[x]_{\mathbf{u}} = {}^t(x_1, \ldots, x_n)$ then

$$q(x) = \sum_{i,j} b_{ij} x_i x_j,$$

which is a quadratic function in the variables $\{x_i\}$. Hence the name.

If the symmetric bilinear pairing is positive definite, i.e. $\langle x, x\rangle \geq 0$ with equality only if $x = 0$, then the quadratic form takes values in $\mathbb{R}_{\geq 0}$ and the value 0 is obtained only for $x = 0$, and vice-versa. We will refer to such quadratic forms as **positive**.

19.2. **Euclidean lattices.** For $v \in \mathbb{R}^n$ and $r \in \mathbb{R}_{\geq 0}$ we denote the open and closed balls of radius r centred at v by

$$B(v, r) = \{u \in \mathbb{R}^n : \|u - v\| < r\}, \quad B[v, r] = \{u \in \mathbb{R}^n : \|u - v\| \leq r\},$$

respectively. If we need to emphasize the ambient space, we shall write $B_n(v, r)$ and $B_n[v, r]$, respectively.

Let us denote $\omega_n = \mathrm{vol}(B_n(0, 1))$ then

$$\omega_1 = 2,\ \omega_2 = \pi,\ \omega_n = \omega_{n-2}\frac{2\pi}{n}. \tag{28}$$

(See (5) for a closed formula.)

Recall that an abelian group A is called **free of rank** n if there are elements $a_1, \ldots, a_n$ in A such that any element of A can be written as $\sum_{i=1}^n n_i a_i$ for some uniquely determined coefficients $n_i \in \mathbb{Z}$; otherwise said $A = \oplus_{i=1}^n \mathbb{Z}a_i$. Equivalently, $A \cong \mathbb{Z}^n$.

A **lattice** $\mathscr{L} \subset \mathbb{R}^n$ is a free abelian group of rank n which is **discrete**: there exists an $\epsilon > 0$ such that

$$\mathscr{L} \cap B(0, \epsilon) = \{0\}.$$

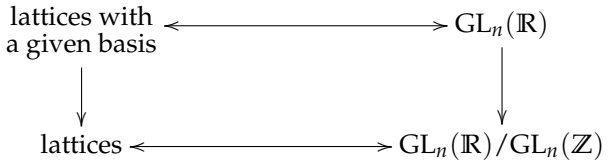
Note the following: suppose that $\mathscr{L} = \oplus_{i=1}^n \mathbb{Z}v_i$. If there exists a basis $\{u_i\}$ for $\mathbb{R}^n$ such that $\{u_i\} \subset \mathscr{L}$ then $\{v_i\}$ is a basis for $\mathbb{R}^n$ too. We emphasize that when we speak of a lattice in $\mathbb{R}^n$

we always assume it has rank n; some authors refer to such a lattice as a "full lattice", but we will not use this terminology.

Exercise 19.2.1. ★ Let $\mathscr{L} \subset \mathbb{R}^n$ be a free abelian group of rank n. Then, $\mathscr{L}$ is a lattice if and only if $\mathscr{L}$ contains a basis of $\mathbb{R}^n$.

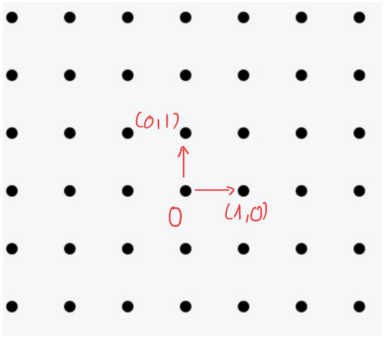
The exercise provides us with a simple method to construct lattices. Let $\{v_1, \dots, v_n\}$ be a basis for $\mathbb{R}^n$, equivalently, the matrix $A = (v_1|v_2|\dots|v_n)$ is in $\mathrm{GL}_n(\mathbb{R})$. Then $\mathscr{L} = \oplus_{i=1}^n \mathbb{Z}v_i$ is a lattice; the matrix A is called a **generator matrix** for $\mathscr{L}$. Conversely, given a lattice $\mathscr{L}$ choose a basis $\{v_1, \dots, v_n\}$ for $\mathscr{L}$ so that $\mathscr{L} = \oplus_{i=1}^n \mathbb{Z}v_i$. Then $(v_1|v_2|\dots|v_n)$ is in $\mathrm{GL}_n(\mathbb{R})$.

If we choose a different basis $\{u_1, \dots, u_n\}$ to $\mathscr{L}$, then there is an invertible matrix $M \in \mathrm{GL}_n(\mathbb{Z})$ such that $(u_1|u_2|\dots|u_n) = (v_1|v_2|\dots|v_n)M$. (This just means that if $M = (m_{ij})$ then $u_1 = m_{11}v_1 + m_{21}v_2 + \dots + m_{n1}v_n$, etc.) And so, instead of the matrix A we get the matrix AM. We conclude the following:

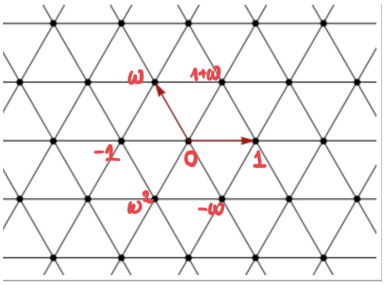


19.2.1. *Examples.* Here are some very basic examples of lattices.

(1) $\mathbb{Z}^n$. We can also think about this lattice as corresponding to $I_n \in \mathrm{GL}_n(\mathbb{R})$, but note that the columns of any matrix $A \in \mathrm{GL}_n(\mathbb{Z})$ also form a basis of $\mathbb{Z}^n$. The Gram matrix associated to I_n is just I_n, but the Gram matrix associated to A is ${}^tAA$, which could be a very complicated matrix. For $n = 3$, for example, we can take $A = \left(\begin{smallmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{smallmatrix}\right)$ to get $B = \left(\begin{smallmatrix} 1 & 2 & 3 \\ 2 & 5 & 8 \\ 3 & 8 & 14 \end{smallmatrix}\right)$.



(2) Identify $\mathbb{C}$ with $\mathbb{R}^2$ so that $a + bi \leftrightarrow (a, b)$, as usual. Let $\omega = e^{2\pi i/3} = \frac{-1+\sqrt{-3}}{2}$ be a third root of unity. Then $1, \omega$ are a basis for $\mathbb{C}$ over $\mathbb{R}$, and the corresponding real vectors $(1, 0)$ and $(-\frac{1}{2}, \frac{\sqrt{3}}{2})$ are basis for $\mathbb{R}^2$. The lattice that we get, which corresponds to the ring $\mathbb{Z}[\omega] = \mathbb{Z} + \mathbb{Z}\omega$, is called the **hexagonal lattice**; it has a generator matrix $\left(\begin{smallmatrix} 1 & -\frac{1}{2} \\ 0 & \frac{\sqrt{3}}{2} \end{smallmatrix}\right)$. (Incidentally, doing the same with the ring of Gaussian integers $\mathbb{Z}[i]$, produces the lattice $\mathbb{Z}^2$.)



Here is another example. The details are left as an exercise.

Exercise 19.2.2. Let $d > 0$ be an integer which is not a square. Consider the ring

$$\mathbb{Z}[\sqrt{d}] = \{a + b\sqrt{d} : a, b \in \mathbb{Z}\}.$$

Prove that the map

$$a + b\sqrt{d} \mapsto (a + b\sqrt{d}, a - b\sqrt{d}) \in \mathbb{R}^2,$$

realizes $\mathbb{Z}[\sqrt{d}]$ as a lattice in $\mathbb{R}^2$. What is the intersection of this lattice with the circle $x^2 + y^2 = 1$? the hyperbola $xy = 1$?

19.3. **Lattices and quadratic forms.** Let $\mathscr{L}$ be a lattice in $\mathbb{R}^n$ and choose a basis $\{v_1, \ldots, v_n\}$ for it. Let $B = (b_{ij}) = (\langle v_i, v_j \rangle)_{ij}$ be the matrix of inner products of the basis vectors. Then B is a symmetric positive definite matrix. Changing the basis amount to changing the matrix B by

$$B \mapsto {}^tMBM, \qquad M \in \mathrm{GL}_2(\mathbb{Z}).$$

Thus, to any lattice there is associated a **similarity class** of positive definite symmetric matrices. A matrix B defined this way is called a **Gram matrix** for the lattice $\mathscr{L}$.

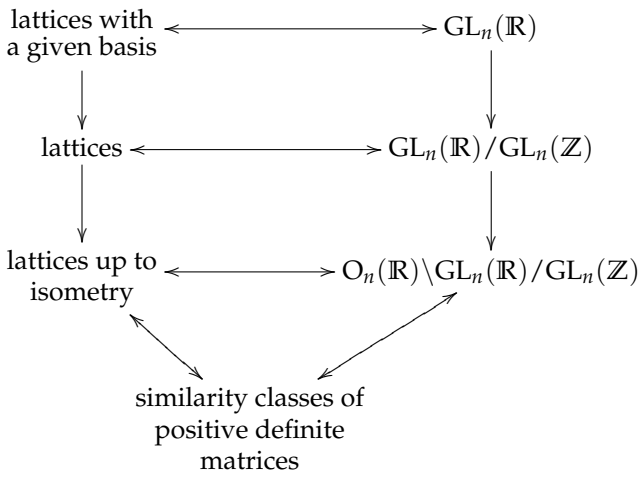
Conversely, given a positive definite symmetric matrix B, we can define a lattice $\mathscr{L}$ such that B is a Gram matrix for $\mathscr{L}$. Indeed, as B is a positive definite symmetric real matrix, there is a matrix $A = (u_1 | \cdots | u_n) \in \mathrm{GL}_n(\mathbb{R})$ such that

$$B = {}^tAA.$$

Let $\mathscr{L}$ be the lattice $\oplus_{i=1}^n \mathbb{Z}u_i$.

Note that there is a little snag, though. The matrix A is unique only up to matrices N such that ${}^tNN = I_n$. Namely, up to an orthogonal matrix $N \in \mathrm{O}_n(\mathbb{R})$. And the lattice associated to NA is usually different than $\mathscr{L}$.

To remedy this we introduce the notion of **isometric lattices**. Two lattices $\mathscr{L}_1, \mathscr{L}_2$ in $\mathbb{R}^n$ are called **isometric** if there is an orthogonal matrix N such that $N\mathscr{L}_1 = \mathscr{L}_2$. Note that if $\mathbf{u} = \{u_i\}$ is a basis for $\mathscr{L}_1$ then $N\mathbf{u} := \{Nu_i\}$ is a basis for $\mathscr{L}_2$ and the Gram matrices for these bases are *equal*. Indeed, $\langle Nu_i, Nu_j \rangle = {}^tu_i{}^tNNu_j = {}^tu_iu_j = \langle u_i, u_j \rangle$. Thus, in fact, we can associate to an isometry class of lattices a similarity class of positive definite matrices $\{{}^tMBM : M \in \mathrm{GL}_n(\mathbb{Z})\}$. And, conversely, to a matrix B we can associate an isometry class of lattices by writing $B = {}^tAA$, as above. We summarize all that by the following diagram:



19.4. **Discriminant, co-volume, and dual lattice.**

19.4.1. Let A be a generator matrix for a lattice $\mathscr{L}$, with columns $u_1, \dots, u_n$. We define the **fundamental parallelepiped** of $\mathscr{L}$ as

$$\mathcal{P} = \{\sum_{i=1}^{n} r_i u_i : 0 \leq r_i \leq 1, i = 1, \dots, n\}.$$

By a well-known property of the determinant,

$$\mathrm{vol}(\mathcal{P}) = |\det(A)|.$$

We define the **co-volume** of $\mathscr{L}$ to be

$$\mathrm{covol}(\mathscr{L}) = \mathrm{vol}(\mathcal{P}) = |\det(A)|.$$

For later use we also define

$$\mathcal{P}^0 = \{\sum_{i=1}^{n} r_i u_i : 0 \leq r_i < 1, i = 1, \dots, n\}.$$

Note that $\mathcal{P}$ and $\mathcal{P}^0$ have the same volume, $\mathcal{P}$ is the closure of $\mathcal{P}^0$, and (as we will prove below)

$$\mathbb{R}^n = \coprod_{\lambda \in \mathscr{L}} \lambda + \mathcal{P}^0.$$

On the other hand, $\mathcal{P}$ and $\lambda + \mathcal{P}$ may intersect, but only along their boundaries.

19.4.2. Let $\mathscr{L} \subset \mathbb{R}^n$ be a lattice. Define its **dual lattice** as

$$\mathscr{L}^\perp := \{v \in \mathbb{R}^n : \langle v, \ell \rangle \in \mathbb{Z}, \forall \ell \in \mathscr{L}\}.$$

In general, the relation $\mathscr{L} \subset \mathscr{L}^\perp$ need not hold; if it does, $\mathscr{L}$ is called an **integral lattice**.

Exercise 19.4.1. Let $\mathscr{L}$ be a lattice with a generator matrix A. Show that $\mathscr{L}$ is integral if and only if its Gram matrix $b = {}^tAA$ has integer entries.

Exercise 19.4.2. Let A be a generator matrix for $\mathscr{L}$. Prove that ${}^tA^{-1}$ is a generator matrix for $\mathscr{L}^\perp$.

Exercise 19.4.3. Let $\mathscr{L}$ be an integral lattice and let $\mathscr{L}_1 \subseteq \mathscr{L}$ be a sub lattice. Prove that $\mathscr{L}_1^\perp \supseteq \mathscr{L}^\perp$ and $[\mathscr{L}_1^\perp : \mathscr{L}^\perp] = [\mathscr{L} : \mathscr{L}_1]$. (Hint: if $[\mathscr{L} : \mathscr{L}_1] = m$ then $\mathrm{covol}(\mathscr{L}_1) = m \cdot \mathrm{covol}(\mathscr{L})$.)

19.4.3. Let $\mathscr{L}$ be a lattice and choose a basis, and thus a Gram matrix B, associated to $\mathscr{L}$. The Gram matrix B is well-defined up to $B \mapsto {}^tMBM$, $M \in \mathrm{GL}_n(\mathbb{Z})$. Note that $\det({}^tMBM) = \det(B)\det({}^tM)\det(M) = \det(B)\det(M)^2 = \det(B)$. Thus, $\det(B)$ is a well defined invariant of the lattice, called its **discriminant**;

$$\mathrm{disc}(\mathscr{L}) = \det(B).$$

Note that

$$\mathrm{disc}(\mathscr{L}) = \mathrm{covol}(\mathscr{L})^2.$$

19.4.4. Suppose now that $\mathscr{L}$ is an integral lattice. Thus, $\mathscr{L} \subseteq \mathscr{L}^\perp$. We have

$$\mathrm{covol}(\mathscr{L}) = [\mathscr{L}^\perp : \mathscr{L}] \times \mathrm{covol}(\mathscr{L}^\perp);$$

on the other hand, using Exercise 19.4.2,

$$\mathrm{covol}(\mathscr{L}) = |\det(A)|, \quad \mathrm{covol}(\mathscr{L}^\perp) = |\det(A^{-1})|.$$

We conclude that

$$\mathrm{disc}(\mathscr{L}) = |\det(A)^2| = [\mathscr{L}^\perp : \mathscr{L}].$$

Exercise 19.4.4. Calculate the discriminant and the dual lattice of the following lattices.

(1) $\mathscr{L} = \mathbb{Z}^n$.
(2) Let m be a positive integer, $\mathscr{L} = \{(a_1, \ldots, a_n) \in \mathbb{Z}^n : \sum_{i=1}^n a_i \equiv 0 \pmod{m}\}$. (The lattices one gets for $m = 2$ are called the D_n lattices.)
(3) $\mathscr{L} =$ the hexagonal lattice.
(4) Let $d > 1$ be a square free integer. Consider the ring $\mathbb{Z}[\sqrt{-d}]$. Under the identification of $\mathbb{C}$ with $\mathbb{R}^2$ it becomes a lattice $\mathscr{L} \subset \mathbb{R}^2$. Write a generator matrix and a Gram matrix for $\mathscr{L}$; find the discriminant and the dual lattice. Is this an integral lattice?

Exercise 19.4.5. ★ A lattice is called **self-dual**, or **unimodular**, if $\mathscr{L} = \mathscr{L}^\perp$. Show that the only unimodular lattice in $\mathbb{R}^2$, up to isometry, is $\mathbb{Z}^2$.

Exercise 19.4.6. Consider the quadratic forms $q(x) = x^2 + y^2$ and $q(x) = x^2 - xy + y^2$. Find lattices in $\mathbb{R}^2$ with these quadratic forms (namely, that they have a Gram matrix with associated quadratic form given by q).

Exercise 19.4.7. Let $(x, y, z) \in \mathbb{R}^3$ and consider the abelian group generated by $(1, 0, 0), (0, 1, 0)$ and (x, y, z). Namely, $\mathbb{Z}(1, 0, 0) + \mathbb{Z}(0, 1, 0) + \mathbb{Z}(x, y, z)$. What are the conditions for it to be free of rank 3? What are the conditions for it to be a lattice? What are the conditions for it to be an integral lattice? a self-dual lattice?

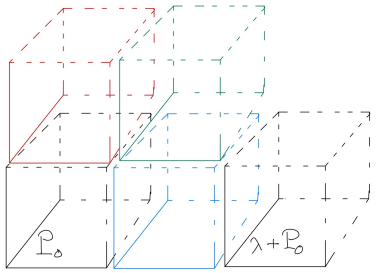
20. Minkowski's Lattice Point Theorem

Let $\mathscr{L}$ be a lattice in $\mathbb{R}^n$, say $\mathscr{L} = \oplus_{i=1}^n \mathbb{Z}v_i$. We have defined above the fundamental parallelepiped $\mathcal{P}$ and the "half open" parallelepiped $\mathcal{P}^0$:

$$\mathcal{P}^0 = \{\sum_{i=1}^n r_i v_i : 0 \le r_i < 1, i = 1, \ldots, n\}.$$

As remarked before, $\mathcal{P}$, which is the closure of $\mathcal{P}^0$ has the same volume as $\mathcal{P}^0$, and

$$\mathbb{R}^n = \coprod_{\lambda \in \mathscr{L}} \lambda + \mathcal{P}^0. \tag{29}$$



Indeed, as $\{v_i\}$ form a basis, any $x \in \mathbb{R}^n$ can be written as $x = x_1v_1 + \cdots + x_nv_n$, with $x_i \in \mathbb{R}$. Then, for $\lambda = \lfloor x_1 \rfloor \cdot v_1 + \cdots + \lfloor x_n \rfloor \cdot v_n \in \mathscr{L}$, we have

$$x \in \lambda + \mathscr{P}^0.$$

Moreover, suppose that $(\lambda_1 + \mathscr{P}^0) \cap (\lambda_2 + \mathscr{P}^0) \neq \emptyset$, then $(\lambda + \mathscr{P}^0) \cap \mathscr{P}^0 \neq \emptyset$, where $\lambda = \lambda_1 - \lambda_2 = a_1v_1 + \cdots + a_nv_n$, for some integers a_i. This implies that for some $0 \leq r_i, s_i < 1$ we have $(a_1 + r_1)v_1 + \cdots + (a_n + r_n)v_n = s_1v_1 + \cdots + s_nv_n$ and so $a_i = s_i - r_i, \forall i$. As $-1 < s_i - r_i < 1$ the only possibility is that $a_i = 0$, and so $\lambda = 0$. Therefore, the claim in (29) holds true.

A set $S \subset \mathbb{R}^n$ is **convex** if $x, y \in S \Rightarrow \alpha x + (1-\alpha)y \in S$ for any $0 \leq \alpha \leq 1$. Namely, if the line segment between any two points of the set is contained in it. A set $S \subset \mathbb{R}^n$ is **centrally symmetric** if $x \in S \Rightarrow -x \in S$. Note that if S is convex and centrally symmetric then $x, y \in S \implies \frac{1}{2}x - \frac{1}{2}y \in S$. We will use this in the proof of Minkowski's theorem.

Theorem 20.0.1 (H. Minkowksi)**.** *Let $\mathscr{L} \subset \mathbb{R}^n$ be a lattice and let K be a centrally symmetric convex bounded set*[36]*. If*

$$vol(K) > 2^n \cdot \mathrm{covol}(\mathscr{L}),$$

then

$$\exists \lambda \in K \cap \mathscr{L}, \quad \lambda \neq 0.$$

Remark 20.0.2. As $\mathscr{L}$ is discrete, it is easy to conclude that if K is closed and bounded, we can replace the strict inequality by $\mathrm{vol}(K) \geq 2^n \mathrm{covol}(\mathscr{L})$ and the theorem still holds true.

Proof. Suppose that this is not the case. Namely, that for all $\lambda \neq 0$ in $\mathscr{L}$, one has $\lambda \notin K$. Let

$$\kappa := ½ \cdot K = \{½ \cdot x : x \in K\}.$$

Then, for all $\lambda \neq 0$, $\lambda \in \mathscr{L}$, we have

$$(\lambda + \kappa) \cap \kappa = \emptyset. \tag{30}$$

Indeed, otherwise $\exists x, y \in K$ such that $\lambda + \frac{1}{2}y = \frac{1}{2}x$ and that implies $\lambda = \frac{1}{2}x - \frac{1}{2}y$, which is a vector in K, since K is centrally symmetric and convex. Contradiction.

We claim that

$$\mathrm{vol}(\kappa) \leq \mathrm{covol}(\mathscr{L}),$$

and this is a contradiction since $\mathrm{vol}(\kappa) = \frac{1}{2^n}\mathrm{vol}(K)$.

To verify the claim note that

$$\kappa = \coprod_{\lambda \in \mathscr{L}} \kappa \cap (\lambda + \mathscr{P}^0) = \coprod_{\lambda_1, \ldots, \lambda_d} \kappa \cap (\lambda_i + \mathscr{P}^0), \tag{31}$$

for some $\lambda_1, \ldots, \lambda_d \in \mathscr{L}$, as κ is bounded.

[36]The boundedness condition is superfluous and can easily be removed.

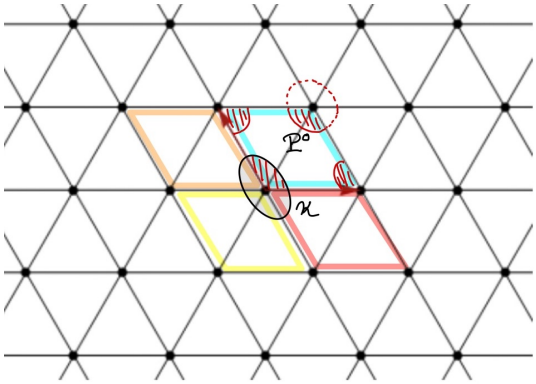
Consider the sets $(\kappa \cap (\lambda_i + \mathscr{P}^0)) - \lambda_i = (\kappa - \lambda_i) \cap \mathcal{P}^0$; they are contained in $\mathcal{P}^0$. We claim that they are disjoint. If not, for some $\lambda_i \neq \lambda_j$, $(\kappa - \lambda_i) \cap (\kappa - \lambda_j) \cap \mathcal{P}^0 \neq \emptyset$, and so $(\kappa - \lambda_i) \cap (\kappa - \lambda_j) \neq \emptyset$. Translating by λ_i we find

$$\kappa \cap (\kappa + (\lambda_i - \lambda_j)) \neq \emptyset,$$

which contradicts (30). Therefore,

$$\begin{aligned}\mathrm{vol}(\kappa) = \sum_{i=1}^{d} \mathrm{vol}(\kappa \cap (\lambda_i + \mathscr{P}^0)) = \sum_{i=1}^{d} \mathrm{vol}(\kappa \cap (\lambda_i + \mathscr{P}^0) - \lambda_i) \\ = \mathrm{vol}(\cup_{i=1}^{d} (\kappa - \lambda_i) \cap \mathcal{P}^0) \leq \mathrm{vol}(\mathcal{P}^0).\end{aligned}$$

□

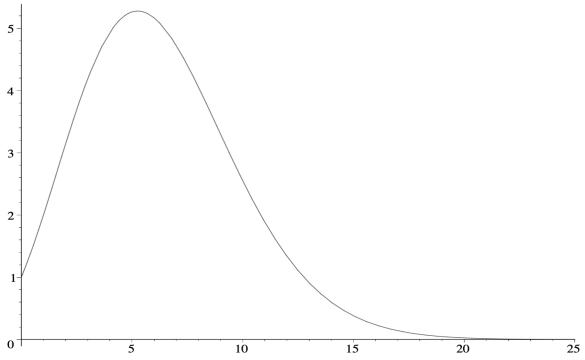


20.1. **Applications of Minkowski's theorem: short vectors.** Let $\mathscr{L} \subset \mathbb{R}^n$ be a lattice. Suppose that $r^n \omega_n \geq 2^n \mathrm{covol}(\mathscr{L})$, where $\omega_n = \mathrm{vol}(B_n(0,1))$. We apply Minkowski's theorem to the closed ball $B_n[0,r]$. Minkowski's theorem implies that there is a $\lambda \in \mathscr{L}$ such that $\lambda \neq 0$ and $\|\lambda\| \leq r$. We deduce the following.

Corollary 20.1.1. *Let $\mathscr{L} \subset \mathbb{R}^n$ be a lattice. $\mathscr{L}$ contains a non-zero vector of length at most*

$$2\sqrt[n]{\frac{\mathrm{covol}(\mathscr{L})}{\omega_n}}.$$

Remark 20.1.2. Note that for $n \gg 0$, ω_n is very small. Thus, dividing by ω_n, increases the quantity under the root. Plotting the volume ω_n as a continuous function of n one finds the following graph:



Exercise 20.1.3. Prove that $\omega_n \geq \left(\frac{2}{\sqrt{n}}\right)^n$ and deduce that $\mathscr{L}$ contains a non-zero vector of length at most

$$\sqrt{n} \cdot (\mathrm{covol}(\mathscr{L}))^{1/n}.$$

20.2. **Applications of Minkowski's theorem: small values of quadratic forms.** Consider a positive definite symmetric bilinear form

$$B(x,y) = {}^t x B y, \quad B \in M_n(\mathbb{R}), \text{symmetric, positive definite.}$$

Write $B = {}^t A A$, $A = (v_1|\dots|v_n) \in \mathrm{GL}_n(\mathbb{R})$ and let $\mathscr{L}$ be the lattice $\mathscr{L} = \oplus_{i=1}^n \mathbb{Z} v_i$. Then

$$\mathrm{covol}(\mathscr{L}) = |\det A| = \det(B)^{1/2}.$$

Applying Corollary 20.1.1, we find that $\mathscr{L}$ contains a vector $x = x_1 v_1 + \dots + x_n v_n$ such that $x_i \in \mathbb{Z}$ and

$$\|x\|^2 \leq 4 \sqrt[n]{\frac{\mathrm{covol}(\mathscr{L})^2}{\omega_n^2}} = 4 \sqrt[n]{\frac{\det(B)}{\omega_n^2}}.$$

But,

$$\|x\|^2 = \langle x, x \rangle = (x_1, \dots, x_n)^t A A \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = B(x,x).$$

Thus, the following theorem follows.

Theorem 20.2.1. *Let B be a symmetric positive-definite real bilinear form on $\mathbb{R}^n$. Then, there is a non-zero vector $x = (x_1, \dots, x_n)$ with integral coordinates such that*

$$B(x,x) \leq 4 \sqrt[n]{\frac{\det(B)}{\omega_n^2}}.$$

20.3. **Applications of Minkowski's theorem: sums of squares.** We apply Corollary 20.1.1 to obtain another proof of a theorem of Fermat that we previously proved using Jacobi sums.

Theorem 20.3.1 (P. de Fermat)**.** *Let $p \equiv 1 \pmod 4$ be a prime. There are integers x, y such that*

$$p = x^2 + y^2.$$

Proof. We construct a sublattice $\mathscr{L}$ of $\mathbb{Z}^2$. Choose $u \in \mathbb{Z}$ such that $u^2 + 1 \equiv 0 \pmod p$. Such exists because $p \equiv 1 \pmod 4$. Let

$$\mathscr{L} = \{(x,y) \in \mathbb{Z}^2 : y \equiv ux \pmod p\}.$$

$\mathscr{L}$ is a lattice and, in fact, has a basis $(1,u), (0,p)$. Therefore, $\mathrm{covol}(\mathscr{L}) = \det \left(\begin{smallmatrix} 1 & 0 \\ u & p \end{smallmatrix}\right) = p$. As $\omega_2 = \pi$, Corollary 20.1.1 tells us that there is a non-zero vector $(x,y) \in \mathscr{L}$ such that

$$x^2 + y^2 = \|(x,y)\|^2 \leq \frac{4p}{\pi} < 2p.$$

But,

$$x^2 + y^2 \equiv (y - ux)(y + ux) \equiv 0 \pmod p,$$

and it follows that $x^2 + y^2 = p$. □

Exercise 20.3.2. Prove that if $p > 2$ is a prime, $p \equiv 1 \pmod 3$, then p is of the form $x^2 + 3y^2$.

Exercise 20.3.3. Prove that if $p > 2$ is a prime, $p \equiv 1 \pmod 8$, then p is of the form $x^2 + 2y^2$.

Theorem 20.3.4 (J. L. Lagrange)**.** *Every positive integer n is a sum of 4 squares. That is, $\exists x, y, z, w \in \mathbb{Z}$ such that*

$$n = x^2 + y^2 + z^2 + w^2.$$

The proof we give uses the Hamilton quaternions. We take some time to discuss this important object.

20.3.1. *The Hamilton quaternions.* To begin with, we define the **Hamilton quaternions** $\mathbb{H}$ as a real vector space of dimension 4, with a basis $1, i, j, k$, where i, j, k are formal symbols. Thus,

$$\mathbb{H} = \mathbb{R} \oplus \mathbb{R}i \oplus \mathbb{R}j \oplus \mathbb{R}k.$$

That is, the elements of $\mathbb{H}$ are formal sums

$$\{a + bi + cj + dk : a, b, c, d \in \mathbb{R}\}.$$

There is a natural real vector space structure on $\mathbb{H}$ (and, in particular, addition). To define mulitplication, it is useful to realize $\mathbb{H}$ as a subset of $M_2(\mathbb{C})$. If we let

$$1 = \left(\begin{smallmatrix} 1 & \\ & 1 \end{smallmatrix}\right), \quad i = \left(\begin{smallmatrix} i & \\ & -i \end{smallmatrix}\right), \quad j = \left(\begin{smallmatrix} & 1 \\ -1 & \end{smallmatrix}\right), \quad k = \left(\begin{smallmatrix} & i \\ i & \end{smallmatrix}\right),$$

then

$$a + bi + cj + dk \quad \longleftrightarrow \quad \begin{pmatrix} a + bi & c + di \\ -c + di & a - bi \end{pmatrix}.$$

This bijection respects the vector space structure and we can endow the quaternions with multiplication using multiplication of matrices. In particular,

$$ij = -ji = k, \quad jk = i, \quad ki = j, \qquad i^2 = j^2 = k^2 = -1.$$

We also gain this way two functions:

$$\mathrm{Tr}\colon \mathbb{H} \to \mathbb{R}, \quad \mathrm{Tr}(a + bi + cj + dk) := \mathrm{Tr} \begin{pmatrix} a + bi & c + di \\ -c + di & a - bi \end{pmatrix} = 2a,$$

and

$$\mathrm{Nm}\colon \mathbb{H} \to \mathbb{R}, \quad \mathrm{Nm}(a + bi + cj + dk) := \det \begin{pmatrix} a + bi & c + di \\ -c + di & a - bi \end{pmatrix} = a^2 + b^2 + c^2 + d^2.$$

It follows that if $z_1, z_2 \in \mathbb{H}$ then

$$\mathrm{Tr}(z_1 + z_2) = \mathrm{Tr}(z_1) + \mathrm{Tr}(z_2), \quad \mathrm{Nm}(z_1 z_2) = \mathrm{Nm}(z_1)\,\mathrm{Nm}(z_2).$$

Exercise 20.3.5. Prove that the map $\mathbb{H} \to \mathbb{H}, z \mapsto z^* := \mathrm{Tr}(z) - z$ is an anti-involution. Namely, it satisfies

$$(z_1 + z_2)^* = z_1^* + z_2^*, \quad (z_1 z_2)^* = z_2^* z_1^*.$$

Prove also that $\mathrm{Nm}(z) = zz^*$. (Suggestion: think in terms of matrices.)

Exercise 20.3.6. Prove that $\mathbb{H}$ is a non-commutative division ring (for any $x \neq 0$ there is a y such that $xy = yx = 1$). One reason this is interesting is that there is no commutative division ring of dimension 4 over $\mathbb{R}$, but here we see that there is a non-commutative division ring.

Consider the subset $\mathbb{Z}[i,j,k] := \mathbb{Z} \oplus \mathbb{Z}i \oplus \mathbb{Z}j \oplus \mathbb{Z}k$; it is easy to see that this is a subring of $\mathbb{H}$. The assertion that an integer n is a sum of 4 squares is *equivalent* to saying that n is a norm of an element of $\mathbb{Z}[i,j,k]$. This will be a key point for the proof of the 4-squares theorem.

Exercise 20.3.7. The **Hurwitz quaternions** is the subset of $\mathbb{H}$ given by

$$\mathbb{Z}\left[i,j,\frac{1+i+j+k}{2}\right] = \left\{a+bi+cj+d\,\frac{1+i+j+k}{2} : a,b,c,d \in \mathbb{Z}\right\}$$
$$= \left\{a+bi+cj+dk \in \mathbb{H} : a,b,c,d \in \mathbb{Z} \text{ or } a,b,c,d \in \mathbb{Z}+\tfrac{1}{2}\right\}.$$

Prove that the Hurwitz quaternions form a subring of $\mathbb{H}$. Prove that Nm is still integer-valued on $\mathbb{Z}\left[i,j,\frac{1+i+j+k}{2}\right]$.

20.3.2. *Proof of the 4-squares theorem.* An integer n is a sum of four squares if and only if n is a norm of a quaternion in $\mathbb{Z}[i,j,k]$. Since the norm is multiplicative, it follows that if both n_1 and n_2 are sums of 4 squares, so is n_1n_2. (This is rather laborious to verify by hand!) Therefore, it is enough to prove that every prime is a sum of 4 squares. We note that $2 = 1^2+1^2+0^2+0^2$ and so we look at a prime $p > 2$.

Claim: There exists integers u, v such that

$$u^2 + v^2 = -1 \pmod{p}.$$

Indeed, mod p, u^2 takes $\frac{p-1}{2}+1$ values, as does $-1-v^2$. As there are only p congruence classes mod p, there must be a value taken by both u^2 and $-1-v^2$. That proves the Claim.

Let $\mathscr{L} \subset \mathbb{Z}^4$ be the lattice[37] spanned by the columns of

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ u & v & p & 0 \\ -v & u & 0 & p \end{pmatrix}.$$

Another presentation of $\mathscr{L}$ is as the column vectors

$$\{(a,b,c,d) \in \mathbb{Z}^4 : c = ua+vb \pmod{p}, d = -va+ub \pmod{p}\}.$$

From the generator matrix we see that $\mathrm{covol}(\mathscr{L}) = p^2$. Since $\omega_4 = \pi^2/2$, we conclude from Corollary 20.1.1 that there are $(a,b,c,d) \in \mathscr{L}$ such that

$$a^2+b^2+c^2+d^2 \le \left(2\sqrt[4]{\frac{p^2}{\pi^2/2}}\right)^2 = \frac{4\sqrt{2}}{\pi}\cdot p < 2p.$$

On the other hand, modulo p,

$$a^2+b^2+c^2+d^2 \equiv a^2+b^2+(ua+vb)^2+(-va+ub)^2$$
$$\equiv (1+u^2+v^2)a^2+(1+u^2+v^2)b^2 \equiv 0.$$

[37] We may keep thinking about $\mathscr{L}$ as contained in $\mathbb{Z}[i,j,k]$, but at this point there's no advantage in doing so. We may simply view $\mathscr{L}$ as a sub lattice of $\mathbb{Z}^4$ in $\mathbb{R}^4$.

Thus, $p|(a^2+b^2+c^2+d^2)$ and it follows that

$$p = a^2+b^2+c^2+d^2.$$

□

20.4. **Applications of Minkowski's theorem: Diophantine approximations.** We will now use Minkowski's theorem, or rather Corollary 20.1.1, to give another proof of Dirichlet's Theorem 4.2.1, slightly reformulated.

Theorem 20.4.1 (Dirichlet). *Let $\theta \in \mathbb{R}$. For every $Q \in \mathbb{N}^+$, there exists integers p, q, not both zero, such that $0 \leq q \leq Q$ and*

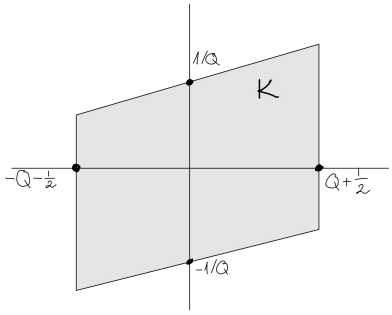
$$|q\theta - p| \leq \frac{1}{Q}.$$

Proof. Let

$$K = \{(x,y) : -Q - \frac{1}{2} \leq x \leq Q + \frac{1}{2},\ |x\theta - y| \leq \frac{1}{Q}\}.$$

We note that K is a convex, centrally symmetric set in $\mathbb{R}^2$ and

$$\mathrm{vol}(K) = (2Q+1)\frac{2}{Q} = 4 + \frac{2}{Q} > 4.$$



For $\mathscr{L} = \mathbb{Z}^2$ we have $\mathrm{covol}(\mathscr{L}) = 1$ and so $\mathrm{vol}(K) > 2^2\mathrm{covol}(\mathscr{L})$. Applying Minkowski's Lattice Point Theorem 20.0.1, we conclude that K contains a non-zero integer vector (q,p); as K is centrally symmetric, we can always arrange that $q \geq 0$. □

Exercise 20.4.2. ★ Prove the following generalization of Dirichlet's theorem, by constructing a suitable convex symmetric set in $\mathbb{R}^{d+1}$. Let $\theta_1, \ldots, \theta_d$ be real numbers and let $Q \in \mathbb{N}^+$. There there are integers $p_1, \ldots, p_d, q$, not all zero, such that $0 \leq q \leq Q$ and

$$|q\theta_i - p_i| \leq \frac{1}{Q^{1/d}},\ \ \forall i.$$

20.5. **Applications of Minkowski's theorem: short solutions to congruences.** Let $z \in \mathbb{Z}^n$, $z = (z_1, \ldots, z_n)$ be a **primitive vector**, i.e. $\gcd(z_1, \ldots, z_n) = 1$. Let $N \geq 2$ be an integer. We are interested in finding non-identically-zero integer solutions to the congruence

$$a_1z_1 + \cdots + a_nz_n = 0 \pmod{N}$$

that are as small as possible in the sense that

$$\|(a_1, \ldots, a_n)\|_\infty := \max\{|a_i| : i = 1, \ldots, n\}$$

is small.

Consider

$$\mathscr{L} = \{(a_1, \dots, a_n) \in \mathbb{Z}^n : \sum a_i z_i \equiv 0 \pmod{N}\}.$$

Note that as $\mathscr{L} \supseteq (N\mathbb{Z})^n$, $\mathscr{L}$ is a lattice. Moreover, its index in $\mathbb{Z}^n$ can be calculated after reduction modulo N,

$$[\mathbb{Z}^n : \mathscr{L}] = [(\mathbb{Z}/N\mathbb{Z})^n : \mathscr{L} \pmod{N}].$$

To calculate this index, we view all vectors as columns vectors. The first fact that we need is left as an exercise:

Exercise 20.5.1. Prove that there is a matrix $M \in \mathrm{GL}_n(\mathbb{Z})$ whose first column is ${}^t(z_1, \dots, z_n)$.

In the notation of the exercise, the condition defining $\mathscr{L}$ modulo N is

$$0 \equiv \sum a_i z_i = (a_1, \dots, a_n)(M^t(1, 0, \dots, 0)) = ((a_1, \dots, a_n)M)^t(1, 0, \dots, 0).$$

That is, M induces a bijection between $\mathscr{L}$ modulo N and the subgroup $\{(0, b_2, \dots, b_n) : b_i \in \mathbb{Z}/N\mathbb{Z}\}$ that has index N in $(\mathbb{Z}/N\mathbb{Z})^n$. Therefore, $[\mathbb{Z}^n : \mathscr{L}] = N$, and it follows that

$$\mathrm{covol}(\mathscr{L}) = N.$$

On the other hand, consider the ball of radius r relative to the norm $\|\cdot\|_\infty$,

$$K(r) = \{(x_1, \dots, x_n) \in \mathbb{R}^n : \max\{|x_i|\} \leq r\}.$$

It is simply the cube $[-r, r]^n$ and has volume $2^n r^n$. If $2^n r^n = 2^n \mathrm{covol}(\mathscr{L})$, that is, if $r = N^{1/n}$, then by Corollary 20.1.1 there is a non-zero vector $(a_1, \dots, a_n) \in \mathscr{L} \cap K(r)$. Namely, a non-zero solution to the congruence mod N such that $|a_i| \leq N^{1/n}$ for all i. We proved the following theorem.

Theorem 20.5.2. *Let $(z_1, \dots, z_n) \in \mathbb{Z}^n$ be a primitive vector. For any $N \in \mathbb{N}^+$ there is a non-zero integer solution $(a_1, \dots, a_n)$ for the congruence*

$$a_1 z_1 + \cdots + a_n z_n \equiv 0 \pmod{N}$$

such that

$$|a_i| \leq N^{1/n}.$$

Example 20.5.3. Suppose that $(z_1, z_2) = (-49, 46)$. We have the solution $(46, 49)$ which is a solution modulo N for every N. Let $N = 31$. Compared to $\sqrt{31} \approx 5.57$ this a big solution. Even noting that this solution modulo 31 is congruent to $(15, -13)$ we get a solution $(15, -13)$ mod 31 that is large compared to $\sqrt{31}$. The theorem says that there is solution (a_1, a_2) modulo 31 with $|a_i| \leq 5$. Indeed, $-1 \times -49 + 5 \times 46 = 279 = 9 \times 31$, so $(-1, 5)$ is a solution modulo 31.

Exercise 20.5.4. Derive a similar theorem for the norm $\|(x_1, \dots, x_n)\|_1 = |x_1| + \cdots + |x_n|$. Namely, in this case we are trying to minimize the total amount of memory needed to store the solution in its entirety and not minimize every x_i separately. Makes sense!

21. Successive minima

21.1. **The shortest vector problem.** Given a lattice $\mathscr{L} \subset \mathbb{R}^n$, what is the shortest non-zero vector in the lattice? Can we find, efficiently, this vector? This problem, which is of theoretical and computational importance, is also currently of great technological importance. The recent activity towards developing post-quantum cryptographic tools (for example, developing procedures for encrypting documents, digitally signing documents, or sharing a secret over open channels) that can withstand attacks by (still putative) quantum computers has put problems

concerning lattices at the forefront; some of the strongest contenders for such procedures are based on hard computational problems concerning lattices, of which the short vector problem (SVP) is an excellent example.

Minkowski's theorem, or rather its direct corollary, Corollary 20.1.1, gave us an estimate: Let $\mathscr{L} \subset \mathbb{R}^n$ be a lattice. $\mathscr{L}$ contains a non-zero vector of length at most

$$2\sqrt[n]{\frac{\operatorname{covol}(\mathscr{L})}{\omega_n}}.$$

While promising, consider the case where $n = 1000$ and the bound for the length is 30, which is not that bad (cf. Exercise 20.1.3). Without further information, to find such a vector we might need to run over all vectors with $0, \pm 1$ coordinates among which 30 or less are not 0, for example; a back of an envelop calculation shows that this requires checking possibly up to 2^{300} different vectors, probably more, which is completely out of the question. Finding a short vector, even if it has relatively small length, is a very hard computational problem! It is believe to be NP hard. That means that by solving this problem one would be able to solve any problem in the complexity class NP, although one cannot determine at this point whether the problem itself is in NP, or perhaps even harder.

One beautiful example of a cryptographic tool using that hard computational problem is a construction due to M. Ajtai and is known as **Ajtai's hash function**. A **hash function** f is a function

$$f\colon \{0,1\}^N \to \{0,1\}^n,$$

where $N \gg n$. Such functions have a variety of applications in cryptography and computer science and we will not get into that here, but provide only one motivation. Our discussion will be very naïve, but it will provide the gist of the role played by hash functions in cryptography. Imagine that you have a very large set of data (grades of students, bank accounts, a genome mapping, ...) and that you want to verify periodically that this data was not compromised, either by intention or due to natural causes. One method is to keep comparing bit-by-bit the file at present time with a recent secure copy of the file. Another option is to apply a hash function to both the secure copy and the current file; a hash function that produces, say, a string of one hundered 0/1 bits. If the result is the same, then, with high probability, the files are identical. Now, based on either statistical arguments, or imagining a scenario with an adversary interested in modifying just a particular part of the data (e.g., adding two zeros to the balance in their bank account), for such applications one wants the hash function to be very sensitive to small changes, one also wants it to be infeasible to reverse-engineer a data file that would hash to a given value.

The following is a variant on Ajtai's hash functions; it is very close to the way these functions appear in the literature. We explain the connection afterwards.

Ajtai's hash function. Fix a reasonably large integer N, say $N = 2^{50}$. Fix a primitive vector

$$(z_1, \ldots, z_n) \in \mathbb{Z}^n,$$

where n is, say, 500. Construct a function

$$h_z\colon \{0,1\}^n \to \mathbb{Z}/N\mathbb{Z}, \quad h_z(\underline{a}) = h_z((a_1, \ldots, a_n)) = \sum_{i=1}^{n} a_i z_i \pmod{N}.$$

If $N = 2^{50}$ any elements of $\mathbb{Z}/N\mathbb{Z}$ can be uniquely written as $\sum_{i=0}^{49} \epsilon_i 2^i, \epsilon_i \in \{0,1\}$ and so we can also view the output of h_z as a vector in $\{0,1\}^{50}$, namely, a string of fifty 0/1 bits.

For cryptographic applications we want this function to be **collision resistant**. Namely, it should be infeasible to find $\underline{a} \neq \underline{b}$ such that $h_z(\underline{a}) = h_z(\underline{b})$.

If $h_z(\underline{a}) = h_z(\underline{b})$ then $h_z(\underline{a} - \underline{b}) \equiv 0 \pmod N$ and note that $\|\underline{a} - \underline{b}\|_\infty = 1$. Thus, if one can find collisions, one can find a very short non-zero solution to the problem

$$\underline{a} \in \mathbb{Z}^n, \qquad a_1 z_1 + \cdots + a_n z_n \equiv 0 \pmod N;$$

in fact, a solution with infinity norm 1.

In the literature one finds the following variant. Let A be an integer matrix with n rows and m columns that is **primitive**. That means that the subgroup $L \subset \mathbb{Z}^n$ spanned by the columns of A is free of rank m and is saturated: if for a non-zero integer m and a non-zero vector $v \in \mathbb{Z}^n$ we have $mv \in L$ then $v \in L$. It is equivalent to the statement that one can complete A to a matrix in $\mathrm{GL}_n(\mathbb{Z})$ (cf. Exercise 20.5.3), and also to the statement that the $\mathbb{Z}^n/L$ is torsion-free (hence, free).

One associates to A a hash function

$$h_A \colon \{0,1\}^n \to (\mathbb{Z}/N\mathbb{Z})^m, \qquad \underline{a} \mapsto \underline{a}A \pmod N.$$

Let us denote the columns of A by $z_1, \dots, z_m$. Using the bijection

$$(\mathbb{Z}/N\mathbb{Z})^m \to \mathbb{Z}/N^m\mathbb{Z}, \qquad (t_1, \dots, t_m) \mapsto t_1 + t_2 N + \cdots + t_m N^{m-1},$$

we find that the vector $h_A(\underline{a}) = (\underline{a} \cdot z_1, \dots, \underline{a} \cdot z_m)$ corresponds to $\underline{a} \cdot z_1 + \underline{a} \cdot z_2 N + \underline{a} \cdot z_m N^{m-1}$ which is just

$$\underline{a} \cdot (z_1, N z_2, \dots, N^{m-1} z_m) \pmod{N^m}.$$

Note that if the original matrix was primitive, this is still a primitive vector. Thus, the hardness of the more general Ajtai hash functions is computationally the same as in the case we first presented.

Ajtai,[38] and then Ajtai-Dwork, Goldreich-Goldwasser-Halevi, and others, have improved more and more on the security of such schemes. In particular, they proved that the ability to find collisions for all the hash functions h_z implies that ability to find, approximately (namely, to a good precision), a short basis for any lattice. This latter problem is known to be NP hard.

21.2. **Minkowksi's theorem on successive minima.** Let $\mathscr{L}$ be a lattice in $\mathbb{R}^n$. Let us introduce notation and denote by $\mu_1(\mathscr{L})$ the length of a shortest non-zero vector in $\mathscr{L}$. That is,

$$\mu_1(\mathscr{L}) = \min\{\|v\| : v \in \mathscr{L}, v \neq 0\}.$$

Note that $\mu_1(\mathscr{L})$ is also the shorted possible distance between two distinct vectors in $\mathscr{L}$.

More generally, for $i = 1, 2, \dots, n$ define the **successive minima** of $\mathscr{L}$ by

$$\mu_i(\mathscr{L}) = \inf\{r : \mathrm{rk}_{\mathbb{R}}(\mathrm{Span}(\mathscr{L} \cap B[0,r])) \geq i\}.$$

In words, $\mu_1(\mathscr{L})$ is the minimal real number r for which we can find a non-zero vector of length r in $\mathscr{L}$; $\mu_2(\mathscr{L})$ is the minimal real number r for which we can find two linearly independent vectors of length $\leq r$ in $\mathscr{L}$; $\mu_3(\mathscr{L})$ is the minimal real number r for which we can find three linearly independent vectors of length $\leq r$ in $\mathscr{L}$, and so on. It is easy to see, using that $\mathscr{L}$ is discrete and the balls are closed, that we can replace "inf" by "min" in the definition.

We make some simple observations:

(1) From the definition,

$$0 < \mu_1(\mathscr{L}) \leq \mu_2(\mathscr{L}) \leq \cdots \leq \mu_n(\mathscr{L}) < \infty.$$

(2) From Minkowski's Theorem (cf. Exercise 20.1.3),

$$\mu_1(\mathscr{L}) \leq \sqrt{n} \cdot \mathrm{covol}(\mathscr{L})^{1/n}.$$

[38] M. Ajtai, Generating Hard Instances of the Short Basis Problem, ICALP 1999: Automata, Languages and Programming pp 1–9, LNCS, volume 1644.

The following theorem, sometimes called Minkowski's Second Theorem, of Minkowsk's Successive Minima Theorem, provides information about the successive minima.

Theorem 21.2.1 (Minkowski)**.** *The successive minima satisfy the following inequalities:*

$$\mathrm{covol}(\mathscr{L})^{1/n} \leq \left(\prod_{i=1}^{n} \mu_i(\mathscr{L})\right)^{1/n} \leq \frac{2}{\omega_n^{1/n}} \cdot \mathrm{covol}(\mathscr{L})^{1/n} \leq \sqrt{n} \cdot \mathrm{covol}(\mathscr{L})^{1/n}.$$

In the proof we will need to use **Hadamard's inequality**, which is the following.

Let A be a $n \times n$ real square matrix with columns $v_1, \ldots, v_n$. Then

$$|\det(A)| \leq \prod_{i=1}^{n} \|v_i\|.$$

To prove this statement, we first note that we can rescale the columns and thus assume that each column is a unit vector. Under these conditions, we need to show that $|\det(A)| \leq 1$. Equivalently, that $\det({}^tAA) \leq 1$. The advantage is that $M = {}^tAA$ is a real symmetric positive definite matrix with diagonal entries 1. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of M; they are real and positive. Then, by the inequality of arithmetic and geometric means, we have

$$\det(M)^{1/n} = (\lambda_1 \cdots \lambda_n)^{1/n} \leq \frac{\lambda_1 + \cdots + \lambda_n}{n} = \frac{1}{n}\mathrm{Tr}(M) = 1,$$

and thus $\det(M) \leq 1$.

We now prove Minkowki's theorem.

Proof. Let $x_1, \ldots, x_n$ be vectors in $\mathscr{L}$ such that $\|x_i\| = \mu_i(\mathscr{L})$. We begin with the lower bound.

On the one hand, for $Q = \{\sum_{i=1}^n \alpha_i x_i : 0 \leq \alpha_i \leq 1\}$ we have

$$\mathrm{vol}(Q) \geq \mathrm{vol}(\mathcal{P}),$$

where $\mathcal{P}$ is a fundamental parallelepiped for $\mathscr{L}$. This is true, because the lattice $\mathscr{L}'$ spanned by $\{x_i\}_{i=1}^n$, which might be a *proper* sublattice of $\mathscr{L}$, satisfies, on the one hand, $\mathrm{covol}(\mathscr{L}') = \mathrm{vol}(Q)$ and, on the other hand, $\mathrm{covol}(\mathscr{L}') = [\mathscr{L} : \mathscr{L}']\mathrm{covol}(\mathscr{L}) = [\mathscr{L} : \mathscr{L}']\mathrm{vol}(\mathcal{P})$. Let A be the generator matrix for $\mathscr{L}'$ with columns $x_1, \ldots, x_n$. Hadamard's inequality states that $\mathrm{vol}(Q) = |\det A| \leq \prod_{i=1}^n \|x_i\|$. But this product is simply $\prod_{i=1}^n \mu_i(\mathscr{L})$.

The upper bound is harder. We first apply the Gram-Schmidt process to the vectors $x_1, \ldots, x_n$ to obtain a basis,

$$\tilde{x}_1, \ldots, \tilde{x}_n,$$

to $\mathbb{R}^n$. This basis has the following properties:

(1) $\{\tilde{x}_1, \ldots, \tilde{x}_n\}$ is an orthonormal set.
(2) $\forall i$, $\mathrm{Span}_{\mathbb{R}}(\tilde{x}_1, \ldots, \tilde{x}_i) = \mathrm{Span}_{\mathbb{R}}(x_1, \ldots, x_i)$.
(3) $\tilde{x}_1 = x_1/\|x_1\|$.

Since $\{\tilde{x}_i\}$ is an orthonormal basis, for any $y \in \mathbb{R}^n$ it holds that

$$y = \sum_{i=1}^{n} \langle y, \tilde{x}_i\rangle \cdot \tilde{x}_i, \quad \|y\|^2 = \sum_{i=1}^{n} \langle y, \tilde{x}_i\rangle^2.$$

Consequently, the set $\{y \in \mathbb{R}^n : \sum_{i=1}^n \langle y, \tilde{x}_i\rangle^2 \leq 1\}$ is just the closed unit ball $B_n[0,1]$. We conclude that the ellipsoid

$$T = \left\{ y \in \mathbb{R}^n : \sum_{i=1}^{n} \left(\frac{\langle y, \tilde{x}_i\rangle}{\mu_i}\right)^2 < 1 \right\}$$

has volume

$$\mathrm{vol}(T) = \mu_1 \cdots \mu_n \cdot \omega_n.$$

Key point. T contains no non-zero vector of $\mathscr{L}$.

Proof. Let $y \in \mathscr{L}, y \neq 0$. Let $1 \leq k \leq n$ be the maximal integer such that

$$\mu_k \leq \|y\|.$$

Hence, if $k < n$, $\|y\| < \mu_{k+1}$.

We claim that $y \in \mathrm{Span}_{\mathbb{R}}(x_1, \ldots, x_k)$. This is clear if $k = n$. For $k < n$, if $y \notin \mathrm{Span}_{\mathbb{R}}(x_1, \ldots, x_k)$ then

$$\mathrm{rk}(\mathrm{Span}_{\mathbb{R}}(B[0, \|y\|] \cap \mathscr{L})) \geq k + 1,$$

and that contradicts the definition of μ_{k+1}. Therefore, $y \in \mathrm{Span}_{\mathbb{R}}(x_1, \ldots, x_k)$ and so

$$y \in \mathrm{Span}_{\mathbb{R}}(\tilde{x}_1, \ldots, \tilde{x}_k).$$

Now,

$$\sum_{i=1}^{n} \left(\frac{\langle y, \tilde{x}_i \rangle}{\mu_i} \right)^2 = \sum_{i=1}^{k} \left(\frac{\langle y, \tilde{x}_i \rangle}{\mu_i} \right)^2 \geq \frac{1}{\mu_k^2} \sum_{i=1}^{k} (\langle y, \tilde{x}_i \rangle)^2 = \frac{1}{\mu_k^2} \|y\|^2 \geq 1,$$

and therefore $y \notin T$. □

The *key point* implies, by Minkowski's theorem, that $\mathrm{vol}(T) \leq 2^n \mathrm{covol}(\mathscr{L})$ and so the inequality,

$$\mu_1 \cdots \mu_n \cdot \left(\frac{2}{\sqrt{n}} \right)^n \leq \mu_1 \cdots \mu_n \cdot \omega_n \leq 2^n \mathrm{covol}(\mathscr{L}),$$

from which the upper bound follows (we have also used the bound on ω_n provided in Exercise 20.1.3). □

Exercise 21.2.2. Find the successive minima and $\mathrm{covol}(\mathscr{L})$ for the following lattices. Write numerically the quantities in Minkowski's lattice point and successive minima theorems.

(1) $\mathscr{L} = \mathbb{Z} \oplus \mathbb{Z}i$ identified with $\mathbb{Z}^2$.
(2) $\mathscr{L} = \mathbb{Z} \oplus \mathbb{Z}\omega$, $\omega = \frac{-1+\sqrt{-3}}{2} \subset \mathbb{C} \cong \mathbb{R}^2$.
(3) $\mathscr{L} = \mathrm{Span}_{\mathbb{Z}}((1,0), (r_1, r_2))$, where r_1, r_2 are non-negative real numbers and $r_2 > 1$. (For μ_2 find only an approximation.)

Exercise 21.2.3. **The D_n lattices.** The D_n lattice in $\mathbb{R}^n$ is defined as

$$D_n = \{(x_1, \ldots, x_n) \in \mathbb{Z}^n : \sum_{i=1}^{n} x_i \equiv 0 \pmod 2\}.$$

Compare Exercise 19.4.4. Find its successive minima. Find also $\sharp\{x \in D_n : \|x\| = \mu_1(D_n)\}$.

22. More examples of lattices

The following lattices are classical.

The lattices A_n. For $n \geq 1$, let

$$A_n = \{(x_0, x_1, \ldots, x_n) \in \mathbb{Z}^{n+1} : x_0 + \cdots + x_n = 0\}$$

A_n is a free abelian group of rank n in $\mathbb{R}^{n+1}$ and a basis for it is provided by

$$M = \begin{pmatrix} -1 & 0 & 0 & \cdots & 0 \\ 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \cdots & -1 \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

A_n is not a lattice per ce, because we always demand that a lattice has full rank in the ambient space. But A_n lies on the hypersurface $H = \{\sum x_i = 0\}$, and by choosing an orthonormal set $y_1, \ldots, y_n$ in H, we can identify H, together with the inner product, with $\mathbb{R}^n$ with the standard inner product by sending $\sum a_i y_i$ to $\sum a_i e_i$, $\{e_i\}$ being the standard basis for $\mathbb{R}^n$. Thus, without specifying this explicitly, we will think about A_n as an n-dimensional lattice.

Exercise 22.0.1. Show that A_2 is identified this way with a lattice that is, up to scaling and perhaps rotation, the hexagonal lattice.

The Gram matrix B of A_n is ${}^tMM$, and a calculation gives

$$B = \begin{pmatrix} 2 & -1 & 0 & \cdots & & 0 \\ -1 & 2 & -1 & \cdots & & 0 \\ 0 & -1 & 2 & \cdots & & 0 \\ \vdots & & & & & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{pmatrix}.$$

Exercise 22.0.2. Prove that $\mathrm{covol}(A_n) = \sqrt{n+1}$. Prove that $\mu_i(A_n) = \sqrt{2}$ for all i. Find

$$\sharp\{x \in A_n : \|x\| = \mu_1\}.$$

(Compare this with the lattice $\mathbb{Z}^n$ that also has all its successive minima equal but for which $\sharp\{x \in \mathbb{Z}^n : \|x\| = \mu_1\} = 2n$.)

Exercise 22.0.3. Is it true or not that A_3, properly rescaled, is isometric to D_3? What about A_4 and D_4?

The lattices A_n^*. This is another notation for the dual lattice $A_n^\perp$, where the dual is calculated in $H \cong \mathbb{R}^n$, and not in $\mathbb{R}^{n+1}$. Otherwise said,

$$A_n^* = \{x \in H : \langle x, y \rangle \in \mathbb{Z}, \forall y \in A_n\}.$$

Exercise 22.0.4. Find a generator matrix for A_n^*. Determine $\mathrm{covol}(A_n^*)$ and prove that $\mu_1(A_n^*) = \sqrt{n/(n+1)}$ and that it is achieved $2n+2$ times if $n \geq 2$ and 2 times if $n = 1$.

The lattice E_6. The lattice E_6 can be constructed as a sublattice of the lattice E_8 we will construct later. But we can also present it explicitly. It lies in the codimension 2 subspace of $\mathbb{R}^8$ given by $\sum x_i = 0$ and $x_1 + x_8 = 0$. A generator matrix is provided by

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & ½ \\ -1 & 0 & 0 & 0 & 0 & ½ \\ 1 & -1 & 0 & 0 & 0 & ½ \\ 0 & 1 & -1 & 0 & 0 & ½ \\ 0 & 0 & 1 & -1 & 0 & -½ \\ 0 & 0 & 0 & 1 & -1 & -½ \\ 0 & 0 & 0 & 0 & 1 & -½ \\ 0 & 0 & 0 & 0 & 0 & -½ \end{pmatrix}$$

Exercise 22.0.5. ★ Write down the Gram matrix and calculate $\mathrm{covol}(E_6)$. Show that $\mu_1(E_6) = \sqrt{2}$ and it is achieved by 72 vectors – this is called the **kissing number** of the lattice. (Note that for $\mathbb{Z}^6$ this number is 12, for A_6 it is 42 and for D_6 it is 60.) The lattice E_6 is known to achieve the highest kissing number among all lattices in $\mathbb{R}^6$, D_4 and D_5 hold the record in their respective dimensions, and A_2 and A_3 in theirs.

23. The sphere packing problem

The **sphere packing** problem ask for the densest packing of solid identical spheres in $\mathbb{R}^n$. Here "sphere" means a closed ball $B_n[v,r]$ of some radius $r > 0$. By a **packing** we means a subset $\mathscr{P}$ in $\mathbb{R}^n$ such that

$$\mathscr{P} = \cup_{i \in I} B_n[v_i, r],$$

over some countable index set I and so that for $i \neq j$, $B_n[v_i, r]$ intersects $B_n[v_j, r]$ at most along their boundaries. Note that

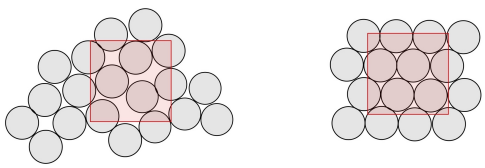
$$\mathrm{vol}(B_n[v_i, r]) = r^n \omega_n,$$

where $\omega_n = \mathrm{vol}(B_n[0,1])$ and is given in Equation (5).

We define the **packing density** as

$$\Delta(\mathscr{P}) = \lim_{N \to \infty} \frac{\mathrm{vol}(\mathscr{P} \cap [-N, N]^n)}{\mathrm{vol}([-N, N]^n)},$$

and we will assume it exists. The sphere packing problem is then to maximize $\Delta(\mathscr{P})$. Note that $\Delta(\mathscr{P})$ is invariant under rescaling; namely, we may assume $r = 1$, but it will be convenient to allow a general r.



23.1. **Lattice packing.** Our main focus will be on **lattice packing**. Let $\mathscr{L} \subset \mathbb{R}^n$ be a lattice and let $r > 0$ such that $\cup_{\lambda \in \mathscr{L}} B[\lambda, r]$ is a packing; it is called a lattice packing. The maximal r such that this is still a packing is called the **packing radius**. We shall denote it $\rho(\mathscr{L})$. Clearly,

$$\rho(\mathscr{L}) = \frac{1}{2}\mu_1(\mathscr{L}).$$

We will denote $\mathscr{P}(\mathscr{L})$ the corresponding packing and by $\Delta(\mathscr{L})$ its density. Another customary notation is the **centre density**,

$$\delta(\mathscr{L}) = \Delta(\mathscr{L})/\omega_n.$$

Lemma 23.1.1. *We have*

$$\Delta(\mathscr{L}) = \frac{\mathrm{vol}(B[0, \frac{1}{2}\mu_1(\mathscr{L})])}{\mathrm{covol}(\mathscr{L})} = \frac{\mu_1(\mathscr{L})^n \omega_n}{2^n \cdot \mathrm{covol}(\mathscr{L})},$$

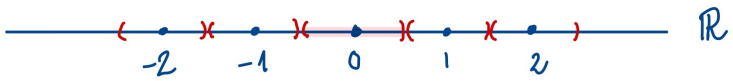
and

$$\delta(\mathscr{L}) = \frac{\mu_1(\mathscr{L})^n}{2^n \cdot \mathrm{covol}(\mathscr{L})}.$$

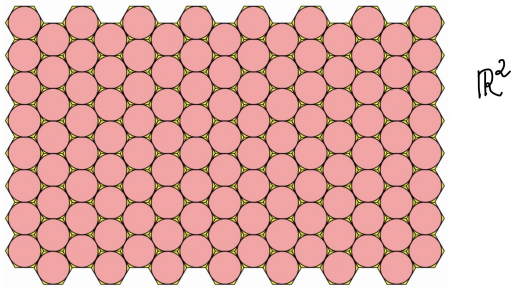
The idea of the proof is quite clear, so we will state it and omit the details. For a very large N, the number of translated fundamental parallelepiped contained in $[-N, N]^n$ is obtained as roughly the ratio of the volumes: $(2N)^n/\mathrm{covol}(\mathscr{L})$. It is also essentially the same number of translated parallelepiped required to cover $[-N, N]^n$ and it is also, roughly, the number of balls in $\mathscr{P}(\mathscr{L})$ contained in $[-N, N]^n$ and the number of balls intersecting non-trivially $[-N, N]^n$.This produces an estimate for $\frac{\mathrm{vol}(\mathscr{P} \cap [-N,N]^n)}{\mathrm{vol}([-N,N]^n)}$ from which N disappears by cancellation and yields by passing to the limit on N the formula in the lemma.

It is not clear *at all* that the best packing densities can always be obtained from lattice packing. Here is the state of the art (February 2021):

(1) $n = 1$. This is true and trivial. In this case $\Delta(\mathbb{Z}) = 1, \delta = \frac{1}{2}$.

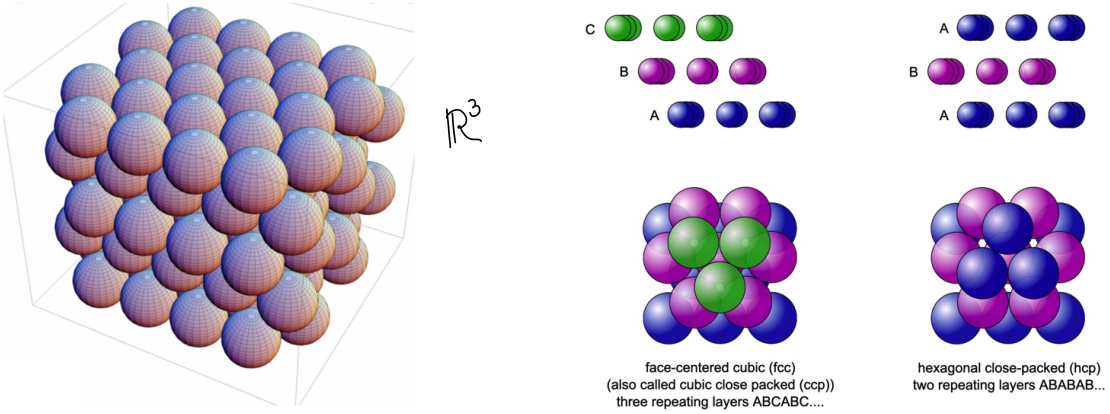


(2) $n = 2$. This is true and non-trivial. The result is due to A. Thue. In this case the lattice is the hexagonal lattice, $\Delta(\mathscr{L}) \approx 0.91, \delta = \frac{1}{2\sqrt{3}}$. That is, the area covered by the discs amounts to about 91% of the total area.



(3) $n = 3$. This is true and extremely hard. The result is due to T. Hales, who gave two proofs. The second one is in collaboration with S. Ferguson and is computer assisted. In fact, there is more than one lattice achieving this density, one being the fcc, or D_3, lattice. There are also non-lattice packing achieving the same density (for example, the hcp packing) but they, too, are formed by laying one layer of the hexagonal lattice on top of itself

The (now proven) conjecture that the best packing is obtained by the fcc lattice is known as **Kepler's conjecture**.



Why is this a difficult problem?? Perhaps the following will give a clue. In the best lattice packing each sphere touches precisely 12 adjacent spheres, but a thirteenth ball very nearly fits. In fact, this problem is known as the kissing problem and it asks how many spheres can touch a given sphere in $\mathbb{R}^n$. In $\mathbb{R}^2$ the number is 6. In $\mathbb{R}^3$ the number is known to be 12, but in Newton's times this was an issue of a controversy, with Newton siding with "12" camp. The solution to the kissing problem is known in dimension 4 (24), 8 (240) and 24 (196,560), but in no other dimensions. The lattices responsible for these kissing numbers are D_4, E_8 and Λ_{24} (see below and Exercise 21.2.3). It is worth noticing that the kissing number of $\mathbb{Z}^4$ is 8, while for D_4 it is 24.

However, the fact that in $\mathbb{R}^3$ one can very nearly fit an additional sphere suggests the idea that by somewhat upsetting a lattice packing one might be able to do better. In high dimensions, already in dimension 10 in fact, there are examples of packings that beat the best currently known lattice packing.

(4) In general one has **Roger's bound**, illustrated in the figure below (taken from Conway & Sloane) for the function $\log_2(\delta) + \frac{1}{96}n(24-n)$. It is valid for any packing, lattice or not.

We notice in dimensions $1, 2, 3, 8$ and 24 very good candidates coming from lattices.

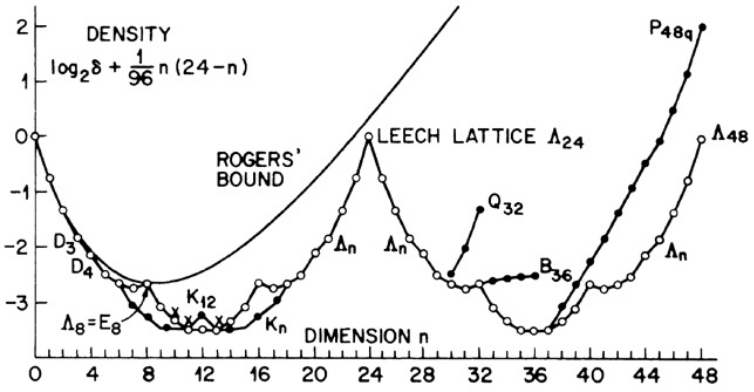
(5) $n = 8$. This is true. Marina Viazovska proved in 2016 that the **E_8-lattice** achieves the best density in $\mathbb{R}^8$. The E_8 lattice has $\Delta \approx 0.25367$, $\delta = 0.0625$.

(6) $n = 24$. This is true. Cohn, Kumar, Miller, Radchenko and Viazovska proved in 2017 that the **Leech lattice** Λ_{24} achieves the best density in $\mathbb{R}^{24}$: $\Delta(\Lambda_{24}) \approx 0.001930$, $\delta(\Lambda_{24}) = 1$. The Leech lattice is named after John Leech who discovered it in 1967[39]

(7) K. Ball proved in 1992 that the best $\Delta(\mathscr{P})$ is at least $2n \cdot 2^{-n}$. This was improved in 2012 by A. Venkatesh:

$$\Delta(\mathscr{P}) \geq \frac{e^{-\gamma}}{2} n \cdot \log\log(n) \cdot 2^{-n}.$$

[39]John Leech, "Notes on sphere packings." Canadian J. Math. 19 (1967), 251–267.



On the other hand, it is known that

$$\Delta(\mathscr{P}) \leq 2^{-0.5990n}.$$

Example 23.1.2. Let us calculate the density of the lattice $\mathbb{Z}^n$. Since $\mu_1(\mathbb{Z}^n) = 1$, we have $\rho(\mathbb{Z}^n) = 1/2$.

$$\Delta(\mathbb{Z}^n) = \frac{\mathrm{vol}(B_n[0,1/2])}{\mathrm{vol}([0,1]^n)} = \frac{\omega_n}{2^n}.$$

Using the estimate $\omega_n \geq 2^n n^{-n/2}$, we get that $\Delta \geq n^{-n/2} = (1/n)^{n/2}$. But note that this is very small compared to the upper bound $2^{-0.5990n} = (1/2^{1.1980})^{n/2}$. (In fact, we can easily prove that there are always lattices with $\Delta \geq (1/4)^{n/2}$; see Proposition 23.1.3.)

So, for example, in dimension 8 we have $\omega_8 = \pi^4/4!$ and $\Delta(\mathbb{Z}^n) = \frac{\pi^4}{2^8 4!}$, while $\Delta(E_8) = \frac{\pi^4}{2^4 4!}$. Namely, the E_8 packing is 16 times more dense than the square packing!

Proposition 23.1.3. *For every n, there is a packing with density $\Delta \geq 2^{-n}$.*

Proof. Consider *any* packing $\mathscr{P}$, lattice or not, by balls of radius r such that one cannot add any ball of radius r to the packing. Without loss of generality $r = 1$. Suppose that

$$\mathscr{P} = \cup_{i\in I} B[v_i, 1].$$

We claim that

$$\mathbb{R}^n = 2\mathscr{P} := \cup_{i\in I} B[v_i, 2].$$

Indeed, if $x \notin B[v_i, 2]$ for any v_i then

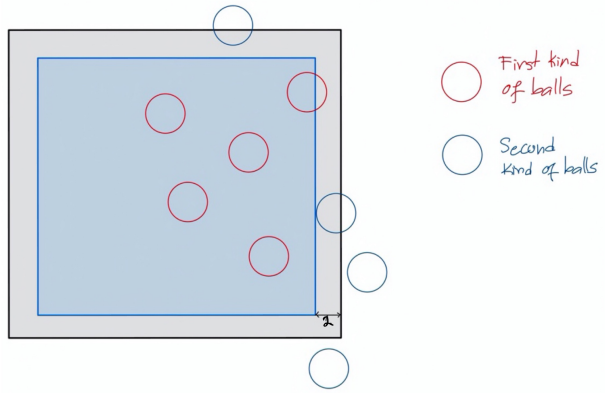
$$\|x - v_i\| > 2, \forall i \in I.$$

That implies that $B[x,1] \cap B[v_i,1] = \emptyset, \forall i \in I$ and so we can add $B[x,1]$ to $\mathscr{P}$. Contradiction.

Now, for every N, $\mathrm{vol}(2\mathscr{P} \cap [-N,N]^n) = \mathrm{vol}([-N,N])^n$. When we compare ball-by-ball the balls $B[v_i,2]$ that contribute to $\mathscr{P} \cap [-N,N]^n$ we can divide them into two groups:

In the first group are those such that $B[v_i,1]$ is entirely contained in $[-N,N]^n$. For such balls $\mathrm{vol}(B[v_i,1] \cap [-N,N]^n) \geq \frac{1}{2^n}\mathrm{vol}(B[v_i,2] \cap [-N,N]^n)$ (with equality only achieved if $B[v_i,2]$ is entirely contained in $[-N,N]^n$, but that doesn't matter to us).

In the second group are the balls $B[v_i,1]$ that aren't contained in $[-N,N]^n$ entirely, but $B[v_i,2]$ intersects non-trivially $[-N,N]^n$. These $B[v_i,2]$ have the property that the distance of v_i from the boundary of $[-N,N]^n$ is at most 2 and so their intersection with $[-N,N]^n$ is contained in the difference of the cubes $[-N,N]^n \setminus [-(N-4),(N-4)]^n$ (one can do better, but this doesn't



matter). Thus, they contribute to the volume at most $2^n(N^n - (N-4)^n)$. We can therefore conclude that

$$\mathrm{vol}(\mathscr{P} \cap [-N,N]^n) \geq \frac{1}{2^n}\left(\mathrm{vol}(2\mathscr{P} \cap [-N,N]^n) - 2^n(N^n - (N-4)^n)\right).$$

To compute the density we need to divide by the volume of $[-N,N]^n$ which is $2^n N^n$. We find that

$$\Delta(\mathscr{P}) \geq \frac{1}{2^n} - \frac{N^n - (N-4)^n}{2^n N^n} \underset{N\to\infty}{\longrightarrow} \frac{1}{2^n}.$$

□

Exercise 23.1.4. For which n, $\Delta(\mathbb{Z}^n) < 2^{-n}$? What explanation is offered by the proof of Proposition 23.1.3?

23.2. **The covering radius.** Let $\mathscr{L} \subset \mathbb{R}^n$ be a lattice. The **covering radius** $R(\mathscr{L})$ of $\mathscr{L}$ is the minimal real number R such that

$$\mathbb{R}^n = \cup_{\lambda\in\mathscr{L}} B[\lambda, R].$$

In fact, this definition makes sense for any subset $\mathscr{L}$ of $\mathbb{R}^n$ if we allow the possibility that $R(\mathscr{L}) = \infty$. It is can be expressed in terms of Voronoi cells.

Let S be a subset of $\mathbb{R}^n$. For every point in $s \in S$ its **Voronoi cell** is the subset

$$V(s) = \{x \in \mathbb{R}^n : \|x - s\| \leq \|x - t\|, \forall t \in S\}.$$

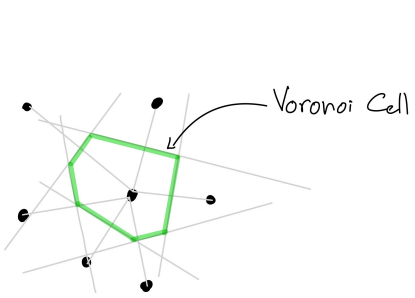
Alternately, for every $t \in S$, $t \neq s$ draw that half space of points closer to s than to t, or of equal distance. It is one of the two half-spaces created by the hyperplane perpendicular to the interval from s to t and passing through its middle point. Then $V(s)$ is the intersection of all these half-spaces. It is a convex closed set and the collection of Voronoi cells covers $\mathbb{R}^n$ with intersections only along their boundaries. If one imagine cell-phone towers positioned at every point of S, the covering radius determines the strength of signal required so that there is full-coverage and is the minimal R such that $\forall s \in S, V(s) \subseteq B[s,R]$.

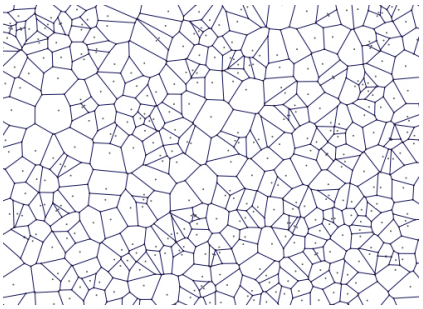
For lattices, the picture that we get is more organized. All the Voronoi cells are polyhedra and are all congruent to each other, each the intersection of finitely many (closed) half-spaces. Let V be the Voronoi cell of 0. We have the following interpretation. The packing radius is

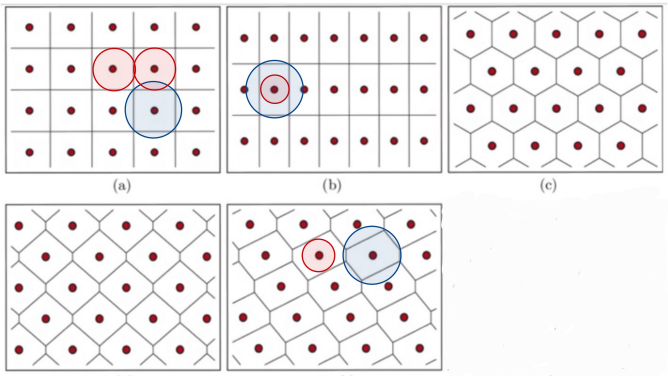
$$\rho(\mathscr{L}) = \text{radius of the largest ball around 0 contained in } V = \frac{1}{2}\mu_1(\mathscr{L}),$$

while the covering radius is

$$R(\mathscr{L}) = \text{radius of the minimal ball around 0 containing } V.$$







Proposition 23.2.1. $R(\mathscr{L}) \geq \frac{1}{2}\mu_n(\mathscr{L})$.

Proof. Suppose that $R = R(\mathscr{L}) < \frac{1}{2}\mu_n(\mathscr{L})$. We construct a set of linearly independent vectors $v_1, \ldots, v_n$ in $\mathscr{L}$ such that

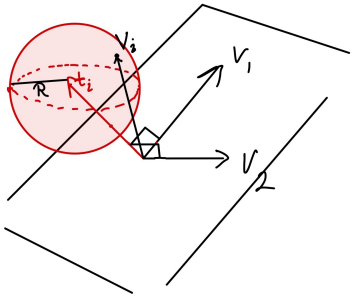
$$\|v_i\| \leq 2R + \epsilon,$$

where $\epsilon = \frac{1}{2}\mu_n(\mathscr{L}) - R > 0$. As $2R + \epsilon = R + \frac{1}{2}\mu_n < \mu_n$ this contradicts the definition of μ_n.

Let $v_0 = 0$ and construct v_i inductively. (One may take v_1 a vector of length $\mu_1(\mathscr{L})$, but the proof works also for $i = 1$.) Assume that $v_1, \ldots, v_{i-1}$ were already defined. Let

$$H_{i-1} = \mathrm{Span}_{\mathbb{R}}(v_1, \ldots, v_{i-1}).$$

Let $t_i \in \mathbb{R}^n$ be a vector perpendicular to H_{i-1} such that $\|t_i\| = R + \epsilon$.



By the definition of R, there is a lattice vector $v_i \in B[t_i, R]$. Note that $v_i \notin H_{i-1}$ (the orthogonal projection of t_i on H_{i-1}, which is of course 0, is also the vector in H_{i-1} closest to t_i and that distance is $R + \epsilon$. As the distance of v_i from t_i is strictly smaller, v_i cannot be in H_{i-1}). It only remains to verify that v_i is short enough:

$$\|v_i\| \leq \|v_i - t_i\| + \|t_i\| \leq R + (R + \epsilon) = 2R + \epsilon.$$

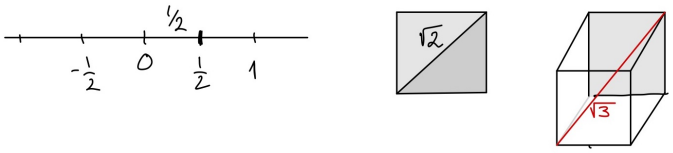
□

Example 23.2.2. The successive minima of $\mathbb{Z}^n$ are $\mu_1(\mathbb{Z}^n) = \cdots = \mu_n(\mathbb{Z}^n) = 1$ as the standard basis vectors are all of length 1. Let us find the covering radius of $\mathbb{Z}^n$. Given $x \in \mathbb{R}^n$ we can find $z \in \mathbb{Z}^n$ such that $|x_i - z_i| \leq 1/2, \forall i$. Indeed, $z_i = \lfloor x_1 \rfloor$ or $\lfloor x_1 \rfloor + 1$. Then, for $z = (z_1, \ldots, z_n)$,

$$\|x - z\| \leq \sqrt{n}/2.$$

For $x = (1/2, \ldots, 1/2)$ we get equality. Hence,

$$R(\mathbb{Z}^n) = \sqrt{n}/2 = \sqrt{n} \times \frac{1}{2}\mu_n(\mathbb{Z}^n).$$



Exercise 23.2.3. ★ Prove that for every lattice $\mathscr{L}$,

$$R(\mathscr{L}) \leq \sqrt{n} \times \frac{1}{2}\mu_n(\mathscr{L}).$$

Exercise 23.2.4. In light of Exercise 23.2.3, $\mathbb{Z}^n$ has the worst covering radius, in the sense that $\frac{R(\mathbb{Z}^n)}{\frac{1}{2}\mu_n(\mathbb{Z}^n)}$ attains the maximum possible. Find this ratio for the hexagonal lattice and the plane lattices $\mathbb{Z}[\sqrt{-d}]$, where $d > 0$ is an integer and we identify $\mathbb{C}$ with $\mathbb{R}^2$.

24. CODES

When we talk about codes, this has nothing to do with secrecy, it has nothing to do with cryptography. Codes are an ingenious device created to maintain integrity of data. Here are typical situations: A rover on Mars transmits pictures to earth over a distance of about 55 million kilometres, when the planets are the closest, using transmission of very limited strength. It is a certainty that some errors will occur along the way. Bits of data may just be lost, or misinterpreted as 0 instead of 1, etc. Errors can also occur at the stage of translating the camera feed into bits for transmission. In another scenario, a scratched DVD is missing some of the data originally recorded on it. Or a computer code may contain errors simply due to the typing process. Codes are used to recognize such errors and recover, with a certain degree of confidence, the original data. In fact, one of the first codes, if not the very first, the Hamming code, was used by R. W. Hamming precisely to correct errors in computer code. The book by T. M. Thompson, *From Error Correcting Codes Through Sphere Packings to Simple Groups*, contains some of the history of the subject.

As this is not a course in Coding Theory, we shall focus on very particular codes. For us, codes will be linear codes over the field $\mathbb{F}_2$ with two elements, denoted $0, 1$. Namely, over $\mathbb{Z}/2\mathbb{Z}$. Such codes are called **binary linear codes**.

24.1. **Codes: first definitions.** A **code** C of **length** n is a subspace of $\mathbb{F}_2^n$. Note that multiplication by scalars is completely degenerate in this case. Thus, a non-empty subset C is a code precisely when it is a subgroup. In fact, precisely when

$$x, y \in C \Longrightarrow x + y \in C.$$

We define the **Hamming distance** on $\mathbb{F}_2^n$ by

$$d(x, y) = \sharp \text{ places where } x \text{ and } y \text{ differ} = \sum_{i=1}^{n} |x_i - y_i|,$$

where $|0| = 0, |1| = |-1| = 1$ (we mention that as, technically, $0, 1 = -1$, are congruence classes and not real numbers so we need to explain what is meant by their absolute value). The **weight** of a vector x is

$$w(x) = d(x, 0) = \sharp \text{ non-zero entries of } x.$$

The **distance** of a code C is

$$d(C) = \min\{w(x) : x \in C, x \neq 0\}.$$

Note that this implies

$$u, v \in C, u \neq v \Longrightarrow d(u, v) \geq d(C).$$

Being a linear subspace, a code has dimension k. This is equivalent to saying that C contains 2^k vectors. Thus, given a code we shall say it is a (n, k, d)**-code**, meaning it is a subspace of $\mathbb{F}_2^n$ (it has length n), has dimension k and distance d.

Exercise 24.1.1. Prove that if C is a (n, k, d)-code then

$$d \leq n - k + 1.$$

Given a code C define the **dual code** $C^\perp$ by

$$C^\perp = \{(y_1, \ldots, y_n) : \sum_{i=1}^{n} x_i y_i = 0, \forall x = (x_1, \ldots, x_n) \in C\}.$$

Namely, under the natural identification of $\mathbb{F}_2^n$ with its dual space (any $(y_1, \ldots, y_n) \in \mathbb{F}_2^n$ defines a linear functional $(x_1, \ldots, x_n) \mapsto \sum_{i=1}^n x_i y_i$ or, more succinctly, $x \mapsto x \cdot y$), $C^\perp$ is just the annihilator of C. Thus, from linear algebra,

$$d(C^\perp) = n - d(C), \qquad (C^\perp)^\perp = C.$$

Let us also define the **Hamming weight enumerator** of C as the polynomial in variables x, y given by

$$W_C(x, y) = \sum_{m=0}^{n} N(m) x^{n-m} y^m,$$

where

$$N(m) = \sharp \{x \in C : w(x) = m\}.$$

Example 24.1.2. Here are some first examples of codes.

(1) The **zero code** Z. This is the code

$$Z = \{0 = (0, \ldots, 0)\}.$$

It is a code of type $(n, 0, 0)$ and

$$W_Z(x, y) = x^n.$$

(2) The **universal code** U. This is the code $Z^\perp$; namely,
$$U = \mathbb{F}_2^n.$$
This is a code of type $(n, n, 1)$ and
$$W_U(x, y) = (x + y)^n = \sum_{m=0}^{n} \binom{n}{m} x^{n-m} y^m.$$

(3) The **repetition code** R. This is the code $\{(0, \dots, 0), (1, \dots, 1)\}$. It is a code of type $(n, 1, n)$ and
$$W_R(x, y) = x^n + y^n.$$

(4) The **parity check code** P. This is the code $R^\perp$:
$$P = \{(x_1, \dots, x_n) : \sum_{i=1}^{n} x_i = 0\} = \{(x_1, \dots, x_n) : x_n = \sum_{i=1}^{n-1} x_i\}.$$
(So, x_n is the parity of the sum of the first $n-1$ digits.) It is a code of type $(n, n-1, 2)$ for $n \geq 2$ and
$$W_P(x, y) = \sum_{m=0,\ m \text{ even}}^{n} \binom{n}{m} x^{n-m} y^m = \frac{1}{2}((x + y)^n + (x - y)^n).$$

Given a code $C \subseteq \mathbb{F}_2^n$ define the **extended code** $C^e \subseteq \mathbb{F}_2^{n+1}$ by
$$C^e = \{(x_1, \dots, x_n, \sum_{i=1}^{n} x_i) : (x_1, \dots, x_n) \in C\}.$$
Thus, the parity check code P of length $n + 1$ is U^e, where U is the universal code in $\mathbb{F}_2^n$.

Exercise 24.1.3. Determine $(n(C^e), k(C^e), d(C^e))$ in terms of $(n(C), k(C), d(C))$. Determine W_{C^e} in terms of W_C. Determine $(C^e)^\perp$ in terms of $C^\perp$.

24.2. **How are codes used?** Codes are used to communicate over noisy channels. Imagine for example a rover on Mars sending data back to earth. This data is a string of zero's and one's.

Space and the atmosphere are full of interference, the signal is weak and is sent over many millions of kilometres. Errors are bound to happen. Let C be an error correcting code of type (n, k, d). Break the original data into blocks of size k and choose a linear isomorphism
$$T : \mathbb{F}_2^k \to C.$$

The transformation T encodes any block of length k as a code word in C. If A is an $n \times k$ matrix whose columns generate C then T is simply $T(x) = Ax$. Now the blocks of length n generated this way are transmitted. Only code words are transmitted in this process, and at the receiving-end blocks of length n are received such that each of them *should* be a code word. This is the case if no errors occurred. But this is not always the case.

If a received vector v is not in the code, one looks for the code word closest to it. This is a well-defined notion as long as at most $\lfloor (d-1)/2 \rfloor$ errors occur, because then there is unique code word closest to v and we choose it as the correction of v. However, we should also ask ourselves how many errors can be detected? Well, if the original code word was v_0 then, even if $d-1$ errors had occurred between transmission and reception of v_0, we can detect that an error had occured. We therefore say that a (n,k,d) code can detect $d-1$ errors and correct up to $\lfloor (d-1)/2 \rfloor$ errors.

After the stage of detection and repair of errors is completed, the received data is decoded using $T^{-1}\colon C \to \mathbb{F}_2^k$ and the original data is thus obtained.

Example 24.2.1. Let us use the code C generated by the following vectors

$$v_1 = (1,1,0,1,0,0,0)$$
$$v_2 = (0,1,1,0,1,0,0)$$
$$v_3 = (0,0,1,1,0,1,0)$$
$$v_4 = (0,0,0,1,1,0,1)$$

This is the $(7,4,3)$ Hamming code to be discussed later. The matrix A is

$$A = \begin{pmatrix} 1&0&0&0 \\ 1&1&0&0 \\ 0&1&1&0 \\ 1&0&1&1 \\ 0&1&0&1 \\ 0&0&1&0 \\ 0&0&0&1 \end{pmatrix}$$

If the data we want to send is $0,0,1,0,1,1,1,1,0,0,1,1,0$ we send instead

$$A\,{}^t(0,0,1,0) = {}^t(0,0,1,1,0,1,0), \quad A\,{}^t(1,1,1,1) = {}^t(1,0,0,1,0,1,1), \quad A\,{}^t(0,1,1,0) = {}^t(0,1,0,1,1,1,0).$$

Suppose that the first word *received* was $u = (1,0,1,1,0,1,0)$. It is not in the code, but we find the vector $(0,0,1,1,0,1,0)$, which is in the code, in distance 1 from u. Thus, it is most likely that u was originally $(0,0,1,1,0,1,0)$.

24.3. **MacWilliams' identity.** Our next result is a beautiful identity between the weight enumerator of a code and its dual. To quote Neil Sloane, "This theorem, due to Mrs. F. J. MacWilliams, is one of the most remarkable results in coding theory".

Theorem 24.3.1 (F. J. MacWilliams' identity)**.** *Let C be a code of dimension k. Then*

$$W_{C^\perp}(x,y) = \frac{1}{2^k} W_C(x+y, x-y).$$

Proof. Given any function

$$f\colon \mathbb{F}_2^n \to R,$$

where $R \supset \mathbb{Q}$ is a commutative ring, define its **Hadamard transform**,

$$\hat{f}(u) = \sum_{v \in \mathbb{F}_2^n} (-1)^{u \cdot v} f(v).$$

The function $\hat{f}$ is again a function $\mathbb{F}_2^n \to R$. The construction is a special case of a Fourier transform and the following formula is a special case of Plancherel's identity, but we will give direct

arguments instead of alluding to Fourier analysis for groups. We claim that

$$\sum_{u\in C^\perp} f(u) = \frac{1}{2^k}\sum_{u\in C}\hat{f}(u). \tag{32}$$

Indeed,

$$\sum_{u\in C}\hat{f}(u) = \sum_{u\in C}\sum_{v\in\mathbb{F}_2^n}(-1)^{u\cdot v}f(v) = \sum_{v\in\mathbb{F}_2^n} f(v)(\sum_{u\in C}(-1)^{u\cdot v}).$$

Now, if $v\notin C^\perp$, $\sum_{u\in C}(-1)^{u\cdot v}=0$ (it is the sum of a non-trivial character $u\mapsto(-1)^{u\cdot v}$, valued in $\{\pm1\}$, on the abelian group C and the usual trick proves the claim). Thus, the last sum reduces to

$$\sum_{v\in C^\perp} f(v)(\sum_{u\in C}(-1)^{u\cdot v}) = 2^k\sum_{v\in C^\perp} f(v),$$

which proves (32).

Note that

$$W_{C^\perp}(x,y) = \sum_{u\in C^\perp} f(u), \quad \text{where } f(u) = x^{n-w(u)}y^{w(u)}\in\mathbb{Q}[x,y].$$

We therefore want to apply the formula (32) to the function $f(u)$. We apply the Hadamard transform to f and find

$$\begin{aligned}\hat{f}(u) = \sum_{v\in\mathbb{F}_2^n}(-1)^{u\cdot v}x^{n-w(v)}y^{w(v)} &= \sum_{v_1=0}^{1}\sum_{v_2=0}^{1}\cdots\sum_{v_n=1}^{1}(-1)^{u\cdot v}x^{n-w(v)}y^{w(v)}\\ &= \sum_{v_1=0}^{1}\sum_{v_2=0}^{1}\cdots\sum_{v_n=1}^{1}\prod_{i=1}^{n}(-1)^{u_i\cdot v_i}x^{1-v_i}y^{v_i} = \prod_{i=1}^{n}\sum_{t=0}^{1}(-1)^{u_i t}x^{1-t}y^t.\end{aligned}$$

Now, if $u_i=0$, the inner sum is $x+y$, while if $u_i=1$ it is $x-y$. Taking now the product, we find that

$$\hat{f}(u) = (x+y)^{n-w(u)}(x-y)^{w(u)} = f(u)(x+y,x-y).$$

Consequently,

$$W_{C^\perp}(x,y) = \sum_{u\in C^\perp} f(u) = \frac{1}{2^k}\sum_{u\in C}\hat{f}(u) = \frac{1}{2^k}W_C(x+y,x-y).$$

□

Exercise 24.3.2. A code is called **self-dual** if $C = C^\perp$. Prove that in this case n is even and $k(C) = n/2$. Prove also that every code word has even weight. Prove that

$$W_C(x,y) = W_C(y,x).$$

(This can be proven using the MacWilliams identity.)

Exercise 24.3.3. Prove that for a self-dual code C,

$$W_C(x,y) = W_C(\frac{x+y}{\sqrt{2}},\frac{x-y}{\sqrt{2}}).$$

Let D be the group of matrices generated by

$$\frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}, \qquad \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}.$$

Prove that D is the dihedral group of 16 elements. Prove that if C is a self-dual code then W_C is invariant under the group D that acts on polynomials by $f(x,y)\mapsto f((x,y)A)$, $A\in D$.

Prove that the polynomials

$$\phi_2 = x^2 + y^2, \quad \phi_8 = x^8 + 14x^4y^4 + y^8,$$

are invariant under D (we will see later that they actually arise from self-dual codes). It is a theorem of A. M. Gleason that for a self-dual code C, W_C is always a polynomial expression in ϕ_2 and ϕ_8. [40]

Exercise 24.3.4. Let C_1, C_2 be codes. The code $C_1 \oplus C_2$ is defined as

$$\{(x, y) : x \in C_1, y \in C_2\}.$$

If C_i is an (n_i, k_i, d_i) code, what is the type of $C_1 \oplus C_2$? Prove that if C_i are both self-dual so is $C_1 \oplus C_2$. Prove that

$$W_{C_1 \oplus C_2}(x, y) = W_{C_1}(x, y) W_{C_2}(x, y).$$

Find all self-dual codes of dimension $2, 4, 6$ and their weight enumerator polynomials. How do your examples compare with Gleason's theorem?

24.4. **Cyclic codes.** Cyclic codes are a simple method to construct codes. In spite of its simplicity, this method produces surprisingly useful codes.

A code C of length n is called **cyclic** if

$$(u_0, u_1, \ldots, u_{n-1}) \in C \Longrightarrow (u_{n-1}, u_0, \ldots, u_{n-2}) \in C.$$

Given a vector $u = (u_0, u_1, \ldots, u_{n-1})$ associate to it a polynomial

$$g_u(t) = u_0 + u_1 t + \cdots + u_{n-1} t^{n-1} \in \mathbb{F}_2[t].$$

Example 24.4.1. Let $C = (\mathbb{F}_2^2)^e = \{(0,0,0), (0,1,1), (1,0,1), (1,1,0)\}$. We see that C is a cyclic code and the polynomials associated to C are

$$0, \quad t + t^2, \quad 1 + t^2, \quad 1 + t.$$

Proposition 24.4.2. *Consider codes of length n.*

(1) There is a bijection

$$\{\text{cyclic codes in } \mathbb{F}_2^n\} \quad \longleftrightarrow \{\text{ideals of the ring } \mathbb{F}_2[t]/(t^n - 1)\}.$$

(2) Any ideal of $\mathbb{F}_2[t]/(t^n - 1)$ is generated by a unique polynomial $g(t) | (t^n - 1)$, $g(t) \in \mathbb{F}_2[t]$.

(3) The dimension of the code corresponding to $g(t)$ is $k = n - \deg(g)$. This code has a basis $\{g(t), tg(t), \ldots, t^{n-\deg(g)-1} g(t)\}$.

Proof. We define

$$J(C) = \{g_u(t) : u \in C\}.$$

We claim that $J(C)$ is an ideal of $\mathbb{F}_2[t]/(t^n - 1)$. First, as $g_{u_1+u_2} = g_{u_1} + g_{u_2}$, $J(C)$ is an abelian group. To prove it is an ideal, it is enough to prove that $tg_u(t) \in J(C)$ for any $u \in C$. If $u = (u_0, u_1, \ldots, u_{n-1})$, then in the ring $\mathbb{F}_2[t]/(t^n - 1)$ we have

$$tg_u(t) = u_0 t + u_1 t^t + \cdots + u_{n-1} t^n = u_{n-1} + u_0 t + u_1 t^2 + \cdots + u_{n-2} t^{n-1} = g_{u'}, \tag{33}$$

where $u' = (u_{n-1}, u_0, \ldots, u_{n-2})$. The vector u' belongs to C because C is cyclic.

[40] A way to prove this result is to calculate the dimension of polynomials of given degree d that are spanned by polynomial expressions in ϕ_2 and ϕ_8, and on the other hand, calculate the dimension of the invariant polynomials of degree d using representation theory. The paper Sloane, N. J. A. Error-correcting codes and invariant theory: new applications of a nineteenth-century technique. Amer. Math. Monthly 84 (1977), no. 2, 82–107, is a nice introduction to these ideas, although it doesn't deal with the exact same result we need here.

Conversely, given an ideal J of $\mathbb{F}_2[t]/(t^n-1)$ associate to it the subset C of $\mathbb{F}_2^n$ given by

$$C = \{u : g_u(t) \in J\}.$$

Since $u_1, u_2 \in J$ implies $u_1 + u_2 \in J$ and $g_{u_1+u_2} = g_{u_1} + g_{u_2}$, C is an abelian group. The computation (33) shows C is cyclic.

The correspondences above are clearly inverses of each other. Note that, in fact, $C \cong J(C)$ as abelian groups. This finishes the proof of the first claim.

To prove (2), we first note the surjective ring homomorphism

$$\mathbb{F}_2[t] \to \mathbb{F}_2[t]/(t^n-1).$$

The surjectivity implies that ideals of $\mathbb{F}_2[t]/(t^n-1)$ correspond bijectively to ideals of $\mathbb{F}_2[t]$ that contain (t^n-1). But, any ideal of $\mathbb{F}_2[t]$ is principal and so generated by a unique polynomial $g(t)$ (usually, $g(t)$ is determined up to a non-zero scalar, but for $\mathbb{F}_2$ this means $g(t)$ is uniquely determined). The ideal $(g(t))$ contains (t^n-1) if and only if $g(t)|(t^n-1)$.

We now consider claim (3). Under the identification

$$C \leftrightarrow J(C), \quad u \leftrightarrow g_u,$$

the code C corresponds to $J(C)/(t^n-1)$ in the ring $\mathbb{F}_2[t]/(t^n-1)$. Since for every polynomial $f(t)$, $\dim_{\mathbb{F}_2}(\mathbb{F}_2[t]/(f(t))) = \deg(f)$, we find that if $J(C) = (g(t))$,

$$\dim_{\mathbb{F}_2}(C) = \dim_{\mathbb{F}_2}(J(C)/(t^n-1)) = n - \deg(g(t)).$$

Let $f(t) = (t^n-1)/g(t)$. There is an isomorphism of vector spaces

$$\mathbb{F}_2[t]/(f(t)) \to J(C)/(t^n-1), \quad h(t) \mapsto h(t)g(t) \pmod{t^n-1}.$$

The basis $1, t, \dots, t^{\deg(f)-1} (= t^{n-\deg(g)-1})$ of $\mathbb{F}_2[t]/(f(t))$ is thus mapped bijectively to the basis $\{g(t), tg(t), \dots, t^{n-\deg(g)-1}g(t)\}$ of $J(C)/(t^n-1)$, that is, of C. □

Example 24.4.3. Some of the codes we have already seen are cyclic codes. For example:

(1) The zero code Z corresponds to the ideal (t^n-1).
(2) The universal code U corresponds to the ideal (1).
(3) The repetition code R corresponds to the ideal $(g(t))$ where $g(t) = (t^n-1)/(t-1)$.
(4) The parity check code P corresponds to the ideal $(t-1)$.

24.4.1. *The Hamming code.* Our next example is one of the most important codes, certainly historically. It is the Hamming code $\mathscr{H}_7$ and the extended Hamming code $\mathscr{H}_8 = \mathscr{H}_7^e$.

The **Hamming code** it the cyclic code in $\mathbb{F}_2^7$ corresponding to the ideal $(1+t+t^3)$ (note that $(1+t+t^3)(1+t+t^2+t^4) = t^7-1$). It is a code of dimension $4 = 7-3$ and a basis is provided by

$$v_1 = (1,1,0,1,0,0,0)$$
$$v_2 = (0,1,1,0,1,0,0)$$
$$v_3 = (0,0,1,1,0,1,0)$$
$$v_4 = (0,0,0,1,1,0,1)$$

As $w(v_i) = 3$, we find $d(\mathscr{H}_7) \leq 3$. We claim that in fact $d(\mathscr{H}_7) = 3$.

If $\mathscr{H}_7$ has a code word of weight 1, using cyclicity, we may assume that it is

$$(1,0,\dots,0) = \epsilon_1 v_1 + \epsilon_2 v_2 + \epsilon_3 v_3 + \epsilon_4 v_4.$$

The first coordinates forces $\epsilon_1 = 1$ and the second coordinate forces then $\epsilon_2 = 1$ and similarly, considering the third and fourth coordinates, we get $\epsilon_3 = 1, \epsilon_4 = 0$. But then the 5-th coordinate is not 0 and that's a contradiction.

If $\mathscr{H}_7$ has a code word of weight 2, we may assume it is $u_a = (1, 0, \dots, 1, \dots)$, where the other 1 appears at the $a+1$ coordinate. Taking a cyclic shift, we get a vector with 1's in the $a+1$ and $2a+1 \pmod 7$ places, whose sum with u_1 is a vector with 1's in the first and $2a+1$ coordinates; we call it u_{2a}. By the same token, we get vectors $u_{2^i a}$ with 1's in the first and $2^i a + 1$ coordinates. As $1 \leq a \leq 6$, we can always get either $2^i a + 1 = 2$ (when $a = 1, 2, 4$), or $2^i a + 1 = 7$ (when $a = 3, 5, 6$) mod 7. Namely, if $\mathscr{H}_7$ has weight 2 it contains either the vector $(1, 1, 0, \dots, 0)$ or $(1, 0, \dots, 0, 1)$. Since the code $\mathscr{H}_7$ is cyclic, it always contains $(1, 1, 0, \dots, 0)$. Writing this vector as $\epsilon_1 v_1 + \epsilon_2 v_2 + \epsilon_3 v_3 + \epsilon_4 v_4$, and considering the first coordinate first, then the second, and so on, we find $\epsilon_1 = 1, \epsilon_2 = 0, \epsilon_3 = 0, \epsilon_4 = 1$. But then the seventh coordinate is not zero. Contradiction.

In summary, $\mathscr{H}_7$ is a $(7, 4, 3)$ code.

Exercise 24.4.4. By considering the cyclic shifts of v_1 conclude that $\mathscr{H}_7$ has at least 7 code words of weight 3. Find a vector of weight 4 and use it to show that $\mathscr{H}_7$ has at least 7 code words of length 4. Show also that there is a code word of weight 7 (Hint: what polynomial will it correspond to?). Explain that this is enough to conclude that

$$W_{\mathscr{H}_7}(x, y) = x^7 + 7x^4y^3 + 7x^3y^4 + y^7.$$

In particular, deduce this way that the distance of $\mathscr{H}_7$ is 3.

Exercise 24.4.5. Prove that $\mathscr{H}_8 := \mathscr{H}_7^e$ is an $(8, 4, 4)$ self-dual code with

$$W_{\mathscr{H}_8}(x, y) = x^8 + 14x^4y^4 + y^8.$$

Recall that a code C is self-dual if $C = C^\perp$. Such a code is always **even**, namely,

$$w(x) \equiv 0 \mod 2, \ \ \forall x \in C.$$

One way to see that is to note that for every $x \in C$,

$$0 = x \cdot x = \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i = w(x) \pmod 2.$$

A self-dual code C is called **doubly even** (or "type II") if

$$w(x) \equiv 0 \pmod 4, \forall x \in C.$$

Corollary 24.4.6. *The extended Hamming code $\mathscr{H}_8$ is a self-dual doubly-even code.*

Let $v \in \mathbb{F}_2^n$ and let r be an integer. By a **ball of distance** r around v, $B[v, r]$, we mean

$$B[v, r] := \{x \in \mathbb{F}_2^n : d(x, v) \leq r\}.$$

What happens if we put balls of radius 1 around each $v \in \mathscr{H}_7$?

(1) The do not intersect. Indeed, a point t of intersection will show that there exists $u, v \in \mathscr{H}_7, u \neq v$, such that $d(u, v) \leq d(u, x) + d(x, v) \leq 2$, which is a contradiction.
(2) Each ball contains 8 vectors – the centre and the 7 vectors in distance 1 from it.
(3) These 2^4 balls contain together $2^4 \times 8 = 2^7$ distinct vectors.

That is, the Hamming code produces a perfect sphere packing in $\mathbb{F}_2^7$; the union of the balls is precisely $\mathbb{F}_2^7$.

A code C of length n is called a **perfect code** if, for an appropriate r, the union of balls of radius r and centres the codes words in C is a disjoint union equal to $\mathbb{F}_2^n$. Such codes exist only for $r = 0, r = n, r = (n-1)/2$ with n odd, $r = 1$ with $n = 2^m - 1$ for $m \geq 1$, and $r = 3$ with $n = 23$. The first three cases are trivial, corresponding to the universal code, the zero code and the code $\{0, (1, \dots, 1)\}$. The case $r = 1$ is not trivial and the simplest example is the Hamming code $\mathscr{H}_7$. It is not hard to prove that if $r = 1$ then one must have $n = 2^m - 1$ for some $m \geq 1$ and the dimension of the code is then $2^m - m - 1$; such codes always exist and a particular construction for every m is given by the so called Hamming codes H_m (where $\mathscr{H}_7 = H_3$). The last example is the Golay code to be discussed below.

24.4.2. *Cyclic codes and duality.* It is natural to ask if the dual code to a cyclic code is also cyclic. It is easy to see that the answer is yes, just from the defining conditions for a vector to be in the dual code. But in fact there is more precise answer.

Theorem 24.4.7. *Let $C \subseteq \mathbb{F}_2^n$ be the cyclic code associated with $g(t)|(t^n - 1)$. Let*

$$h(t) = \frac{t^n - 1}{g(t)}, \quad f(t) = t^{\deg h} h(1/t).$$

The $C^\perp$ is the cyclic code associated with $f(t)$.

Proof. Let us write

$$g(t) = \sum_{i=0}^{d} g_i t^i, \qquad d = \deg(g(t)) \leq n,$$

and

$$h(t) = \sum_{i=0}^{e} h_i t^i, \qquad e = \deg(h(t)) = n - d.$$

Then

$$f(t) = \sum_{i=0}^{e} h_i t^{e-i}.$$

Denote C' the cyclic code generated by $f(t)$. Since $k(C') = k(C^\perp)$, it is enough to show that $C' \subseteq C^\perp$. For that, it is enough to show that every basis element $t^i f(t)$ of C' is perpendicular to every basis element $t^j g(t)$ of C.

Now, in general, under the bijection $\mathbb{F}_2^n \leftrightarrow \mathbb{F}_2[t]/(t^n - 1)$, under which

$$a = (a_0, \dots, a_{n-1}) \leftrightarrow g_a = \sum_{i=0}^{n-1} a_i t^i,$$

we have that $a \cdot b = a_0 b_0 + a_1 b_1 + \cdots + a_{n-1} b_{n-1}$, which is the coefficient of t^{n-1} in the product

$$(a_0 + a_1 t + \cdots + a_{n-1} t^{n-1})(b_0 t^{n-1} + b_1 t^{n-2} + \cdots + b_{n-1}) = g_a(t) \cdot t^{n-1} g_b(1/t).$$

Note that this is calculated in $\mathbb{F}_2[t]/(t^n - 1)$ and so whenever the product $g_a(t) \cdot t^{n-1} g_b(1/t)$ is divisibly by $t^n - 1$, $a \cdot b = 0$. Therefore, we need to calculate the coefficient of t^{n-1} in the polynomial $t^j g(t) \cdot t^{n-1} t^{-i} f(1/t)$. But this just some power of t times $g(t)h(t) = t^n - 1$. □

Exercise 24.4.8. Prove that a cyclic code C associated to $g(t)$ is self-dual, if and only if (in the notation of Theorem 24.4.7) $g(t) = f(t)$, and necessarily n is even. Prove that if $n = 2r$ then $1 + t^r$ defines a cyclic self-dual code.

Exercise 24.4.9. Find all self-dual cyclic codes of length $2, 4, 6, 8, 10$.

Exercise 24.4.10. ★ Find all self-dual cyclic codes of length 14.

24.5. **The Golay code.** Let α be a root of 1 in $\overline{\mathbb{F}}_2$ of order 23. We claim that $\alpha \in \mathbb{F}_{2^{11}}$ and not to any smaller field. First, $2^{11} - 1 = 2047 = 23 \cdot 89$ and so all roots of unity of order 23 lie in $\mathbb{F}_{2^{11}}$. As $[\mathbb{F}_{2^{11}} : \mathbb{F}_2] = 11$ there are no intermediate extension between these two fields. Thus, the minimal polynomial $g(t)$ of α over $\mathbb{F}_2$ must be of degree 11 and divide $t^{23} - 1$. It accounts for only 11 of the 22 primitive roots of unity of order 23. Any root of unity not among them is a root of an additional polynomial $h(t)$ of degree 11 such that $h(t)$ is irreducible and divides $t^{23} - 1$. It follows that the factorization of $t^{23} - 1$ into irreducible polynomials is

$$t^{23} - 1 = (t - 1)g(t)h(t).$$

On the other hand, one may verify that of two following polynomials divides $t^{23} - 1$. Thus, we conclude that for a suitable choice of α,

$$g(t) = 1 + t + t^5 + t^6 + t^7 + t^9 + t^{11}, \qquad h(t) = 1 + t^2 + t^4 + t^5 + t^6 + t^{10} + t^{11}.$$

The cyclic code defined by $g(t)$ is called the **Golay code** $\mathscr{G}_{23}$. We let $\mathscr{G}_{24} = \mathscr{G}_{23}^e$. The Golay codes are still being used by Voyager I and II in transmitting data back to earth over a distance of about 20 billion kilometres at this time! (March 2021). They are truly remarkable and we shall see some evidence of that.

The following theorem gives some properties of the Golay codes. I don't know a proof that is not massively computational, so we will not provide a proof here.

Theorem 24.5.1. *The Golay codes have the following properties:*

(1) $\mathscr{G}_{23}$ is a $(23, 12, 7)$ code, which is perfect.
(2) $\mathscr{G}_{24}$ is a $(24, 12, 8)$ code, which is self-dual and doubly even.
(3) The weight enumerators of these codes are

$$W_{\mathscr{G}_{23}}(x, y) = x^{23} + 253x^{16}y^7 + 506x^{15}y^8 + 1288x^{12}y^{11} + 1288x^{11}y^{12} + 506x^8y^{15} + 253x^7y^{16} + y^{23},$$

and

$$W_{\mathscr{G}_{24}}(x, y) = x^{24} + 759^{16}y^8 + 2576x^{12}y^{12} + 759x^8y^{16} + y^{24}.$$

Assuming that $\mathscr{G}_{23}$ indeed has distance 7 and so is a $(23, 12, 7)$ code, let us check that it is perfect. First, the number of points in a ball of radius 3 in $\mathbb{F}_2^{23}$ is

$$1 + \binom{23}{1} + \binom{23}{2} + \binom{23}{3} = 2^{11}.$$

The number of vectors in $\mathscr{G}_{23}$ is 2^{12}. As balls of radius 3 with centres in $\mathscr{G}_{23}$ do not intersect ($d(\mathscr{G}_{23}) = 7$), the number of points in their union is $2^{11}2^{12} = 2^{23} = \sharp\mathbb{F}_2^{23}$. This shows that the Golay code $\mathscr{G}_{23}$ is perfect.

Exercise 24.5.2. One thought regarding error-correcting is that we may just send every block of size k twice. Consider this for the Golay code. This idea suggests that instead of using the Golay code which is of length 23, we can use the code C of dimension 24 which is a variant on a repetition code.

$$C = \{(x, x) : x \in \mathbb{F}_2^{12}\} \subset \mathbb{F}_2^{24}.$$

Discuss the advantages and disadvantages of this idea.

Exercise 24.5.3. The Golay code $\mathscr{G}_{23}$ turns out to be also a special case of a quadratic residue code (as is the Hamming code $\mathscr{H}_7$). We don't enter into the general theory of such codes here, but it implies that the Golay code is also the cyclic code generated by

$$f(t) = t + t^2 + t^3 + t^4 + t^6 + t^8 + t^9 + t^{12} + t^{13} + t^{16} + t^{18}.$$

(The meaning of that is that the ideal generated by $f(t)$ in $\mathbb{F}_2[t]/(t^n - 1)$ is the same as the one used to define the Golay code.) It also implies that the Hamming code is also generated by

$$t + t^2 + t^4.$$

25. Lattices and codes

What comes next is a rather astounding connection between codes and lattices. The link, called Construction A, is so simple that one hardly suspects it will yield anything interesting, but, in fact, the opposite is true! This construction is due to Neil J. A. Sloane.

25.1. **Construction A of Sloane.** Let C be a code in $\mathbb{F}_2^n$. Define a lattice

$$L(C) \subseteq \mathbb{Z}^n, \qquad L(C) = \{v \in \mathbb{Z}^n : v \in C \pmod 2\}.$$

This method is called Construction A of Sloane. Note that $L(C) \supseteq (2\mathbb{Z})^n$ and therefore is a lattice. Note also that any lattice L such that $(2\mathbb{Z})^n \subseteq L \subseteq \mathbb{Z}^n$ arrises this way. We define

$$\Lambda(C) = \frac{1}{\sqrt{2}} L(C).$$

Before studying the properties of this construction, we recall the notion of kissing number.

Given a lattice $\mathscr{L} \subset \mathbb{R}^n$ recall that its kissing number $\tau(\mathscr{L})$ to be the number of spheres in the lattice packing that touch the sphere at the origin. Otherwise said

$$\tau(\mathscr{L}) = \sharp \{v \in \mathscr{L} : \|v\| = \mu_1(\mathscr{L})\}.$$

Easy examples are

$$\tau(\mathbb{Z}^2) = 4, \quad \tau(\mathbb{Z}^n) = 2n, \quad \tau(\text{Hexagonal lattice}) = 6, \quad \tau(\text{FCC lattice}) = 12.$$

We will soon see an easy way to describe the FCC lattice.

Theorem 25.1.1. *Let C be an (n, k, d)-code. Then*

1. *$\Lambda(C)$ is a lattice in $\mathbb{R}^n$.*
2. *$\operatorname{disc}(\Lambda(C)) = 2^{n-2k}$.*
3. *$\Lambda(C)^\perp = \Lambda(C^\perp)$.*
4. *$\Lambda(C)$ is a self-dual lattice of type II (i.e., for all $x \in \Lambda(C)$, $\|x\|^2 \equiv 0 \pmod 2$) if and only if C is a self-dual code of type II.*
5. *Assume that C is not the zero code. For τ denoting the kissing number, ρ denoting the packing radius and $N(m)$ denoting the number of code words of weight m, we have*

$$\tau(\Lambda(C)) = \begin{cases} 2^d N(d) & d < 4 \\ 2n + 16N(4) & d = 4 \\ 2n & d > 4 \end{cases}, \qquad \rho(\Lambda(C)) = \begin{cases} \frac{1}{2}\sqrt{\frac{d}{2}} & d < 4 \\ \frac{\sqrt{2}}{2} & d = 4 \\ \frac{\sqrt{2}}{2} & d > 4 \end{cases}.$$

Exercise 25.1.2. Verify Theorem 25.1.1 for the codes Z, U, R, P and E_8.

Before proving the theorem, we provide some examples that will convince the reader of its interest.

25.1.1. *Examples of Construction A.*

(1) For the zero code Z, which is an $(n,0,0)$-code, we have

$$L(Z) = (2\mathbb{Z})^n = 2\cdot\mathbb{Z}^n, \qquad \Lambda(Z) = \sqrt{2}\cdot\mathbb{Z}^n.$$

For the universal code $U = Z^\perp$, which is an $(n,n,1)$-code, we have

$$L(U) = \mathbb{Z}^n, \qquad \Lambda(U) = \Lambda(Z)^\perp = \frac{1}{\sqrt{2}}\cdot\mathbb{Z}^n.$$

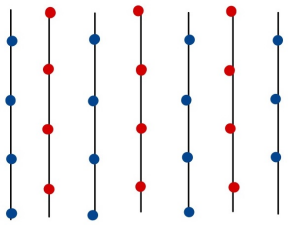
(2) For the repetition code R, which is an $(n,1,n)$-code, we have

$$L(R) = 2\mathbb{Z}^n + \mathbb{Z}(1,1,\ldots,1) = \{x\in\mathbb{Z}^n : x_i \text{ are all even, or are all odd}\}.$$

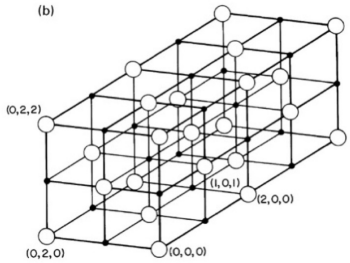
For the parity check code $P = R^\perp$, which is an $(n,n-1,2)$-code, we have

$$L(P) = D^n = \{x\in\mathbb{Z}^n : \sum x_i \equiv 0 \pmod 2\}.$$

For $n=3$, $L(P)$ is the FCC lattice.



L(R), n=2



L(P), n=3

(L(P)=FCC=D_3)

(3) For the extended Hamming code $\mathscr{H}_8$ the construction gives the E_8-lattice

$$\Lambda(\mathscr{H}_8) = E_8.$$

For us, this is the definition of the E_8-lattice, but this lattice appears often in the theory of Lie groups and in physics. It has remarkable symmetry. The group of isometries of E_8 is a group with $696,729,600$ elements, generated by certain reflections in the space $\mathbb{R}^8$. As $\mathscr{H}_8 = \mathscr{H}_7^e$ and we have an explicit basis for $\mathscr{H}_7$ (it is a cyclic code associated to $1+t+t^3$) we easily find that a basis for E_8 is given by the columns of the following matrix:

$$\frac{1}{\sqrt{2}}\times\begin{pmatrix} 1&0&0&0&2&0&0&0\\ 1&1&0&0&0&2&0&0\\ 0&1&1&0&0&0&2&0\\ 1&0&1&1&0&0&0&2\\ 0&1&0&1&0&0&0&0\\ 0&0&1&0&0&0&0&0\\ 0&0&0&1&0&0&0&0\\ 1&1&1&1&0&0&0&0 \end{pmatrix}$$

Recall that E_8 is the best lattice packing, in fact best packing of any sort, in $\mathbb{R}^8$. It is surprising that it has such a simple description. The lattice E_8 turns out to be isometric to the lattice with the following generating matrix - it has the advantage of showing some of the symmetries more clearly:

$$\begin{pmatrix} 2 & -1 & 0 & 0 & 0 & 0 & 0 & 1/2 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 \end{pmatrix}$$

(4) For the extended Golay code $\mathscr{G}_{24} = \mathscr{G}_{23}^e$ we get a self-dual lattice of type II in $\mathbb{R}^{24}$. It is closely related to the Leech lattice – the lattice which optimizes sphere packing, lattice or not, in $\mathbb{R}^{24}$. If we let

$$\Lambda^0 = \{\frac{1}{\sqrt{2}}v : \sum_{i=1}^{24} v_i \equiv 0 \pmod 4\} \subset \Lambda(\mathscr{G}_{24}),$$

then Λ^0 is of index 2 in $\Lambda(\mathscr{G}_{24})$, hence a lattice, and the Leech lattice is

$$\Lambda_{24} := \Lambda^0 + \mathbb{Z}t, \qquad t = \frac{1}{2\sqrt{2}}(-3, 1, 1, \dots, 1).$$

(5) To show to what extent the lattices obtained by Construction A are accessible, we state here a theorem without proof. Let $\mathscr{L} \subset \mathbb{R}^n$ be an integral lattice and define it **theta series** as

$$\Theta_{\mathscr{L}}(q) = \sum_{m=0}^{\infty} r(m)q^m,$$

where

$$r(m) = \sharp\{v \in \mathscr{L} : \|v\|^2 = m\}.$$

Theorem 25.1.3. *Let $\mathscr{L} = \Lambda(C)$, and let W_C be the weight enumerator polynomial of C. Then,*

$$\Theta_{\mathscr{L}}(q) = W_C(\theta_3(q^2), \theta_2(q^2)),$$

where

$$\theta_2(q) = \sum_{m=-\infty}^{\infty} q^{(m+\frac{1}{2})^2}, \quad \theta_3(q) = \sum_{m=-\infty}^{\infty} q^{m^2}.$$

Note that

$$\theta_2(q^2) = 2q^{1/2}(1 + q^4 + q^{12} + q^{24} + \dots)$$

(the exponents are four times triangular numbers: i.e. four times a number of the form $\frac{m(m+1)}{2}$, where m is allowed to be negative). Also,

$$\theta_3(q^2) = 1 + 2(q^2 + q^8 + q^{18} + \dots)$$

where the exponents are twice a square.

Exercise 25.1.4. Prove the identity $\theta_{\mathscr{L}_1 \oplus \mathscr{L}_2} = \theta_{\mathscr{L}_1}\theta_{\mathscr{L}_2}$. Prove that the coefficient of q^m in $(\theta_{\mathbb{Z}})^4$ is positive for every $m \geq 0$.

Exercise 25.1.5. Write an expression for Θ_{E_8} in terms of θ_2 and θ_3. Use it to find the first minimum of E_8 and its kissing number. Using the generator matrix for E_8 now determine all the successive minima of E_8.

25.1.2. *Proof of Theorem 25.1.1.* We begin with the last claim. We may work with $L(C)$ instead of $\Lambda(C)$, properly rescaling everything. It is clear that the vectors of minimal length will have coordinates bounded in absolute value by 2, because we can otherwise add to them a multiple of $2e_i$, where $\{e_i = (0,\dots,\underset{i}{1},\dots,0) : i = 1,\dots,n\}$ are the standard basis, staying in the lattice, but reducing the length.

For $d > 4$, $\mu_1(L(C))$ is achieved by the vectors $\{\pm 2e_i\}$ and there are $2n$ of them. So,

$$\rho(\Lambda(C)) = \frac{1}{2}\cdot\mu_1(\Lambda(C)) = \frac{1}{2}\cdot\frac{2}{\sqrt{2}} = \frac{\sqrt{2}}{2}, \qquad \tau(\Lambda(C)) = 2n.$$

For $d < 4$, since $C \neq Z$, $\mu_1(L(C))$ is achieved, in particular, on the $N(d)$ vectors in C of weight d, when we think about the coordinates of these vectors as integers $0, 1$, thereby viewing them as vectors in $L(C)$. In fact, all the vectors in

$$\{(\pm x_1,\dots,\pm x_n) : x_i \in \{0,1\}, (x_1,\dots,x_n) \in C\}$$

have this same length $\sqrt{d}$. There are $2^d N(d)$ such vectors. Any other vector in $L(C)$ is obtained from one of these by addition or subtraction of multiples of some $2e_i$ and, except for the vectors already taken into account, will have greater length. Therefore,

$$\rho(\Lambda(C)) = \frac{1}{2}\cdot\mu_1(\Lambda(C)) = \frac{1}{2}\cdot\frac{\sqrt{d}}{\sqrt{2}}, \qquad \tau(\Lambda(C)) = 2^d N(d).$$

Finally, for $d = 4$, by similar considerations, $\mu_1(L(C))$ is achieved on the $2^d N(d)$ vectors $\{(\pm x_1,\dots,\pm x_n) : x_i \in \{0,1\}, (x_1,\dots,x_n) \in C, w(x) = 4\}$, as well as on the $2n$ vectors $\pm 2e_i$. Thus,

$$\rho(\Lambda(C)) = \frac{1}{2}\cdot\mu_1(\Lambda(C)) = \frac{1}{2}\cdot\frac{2}{\sqrt{2}} = \frac{\sqrt{2}}{2}, \qquad \tau(\Lambda(C)) = 16N(4) + 2n.$$

The proof of (2) is rather simple. We have $\mathrm{disc}(\Lambda_C) = \mathrm{covol}(\Lambda(C))^2 = 2^{-n}\mathrm{covol}(L(C))^2$. But, $(2\mathbb{Z})^n \underset{2^k}{\subseteq} L(C) \underset{2^{n-k}}{\subseteq} \mathbb{Z}^n$, and so $\mathrm{covol}(L(C)) = 2^{n-k}\mathrm{covol}(\mathbb{Z}^n) = 2^{n-k}$ and (2) follows.

For proving claims (3) - (4), we write $\mathbb{F}_2^n$ as columns vectors. Note that the symmetric group S_n acts as isometries of both $\mathbb{F}_2^n$, that is, it preserves the bilinear form $x\cdot y = \sum x_i y_i \pmod 2$ and the Hamming distance; S_n also acts as isometries of $\mathbb{R}^n$, which means that it preserves the bilinear form $x\cdot y = \sum x_i y_i$. Thus, we can permute the coordinates if needed. In particular, taking an $n\times k$ generator matrix for a code C of type (n,k,d), after applying column reduction we may assume that C has a generator matrix of the form

$$g = \begin{pmatrix} I_k \\ B \end{pmatrix} \in M_{n,k}(\mathbb{F}_2),$$

for some matrix B in $M_{n-k,k}(\mathbb{F}_2)$.

Claim. The dual code $C^\perp$ *has a generator matrix*

$$h = \begin{pmatrix} -B^t \\ I_{n-k} \end{pmatrix} \in M_{n,n-k}(\mathbb{F}_2).$$

Proof. As the matrix h above has rank $n-k$, it is enough to prove that the span of its columns is contained in $C^\perp$. Namely, it is enough to verify that ${}^t hg = 0$. Indeed,

$${}^t hg = \begin{pmatrix} -B & I_{n-k} \end{pmatrix}\begin{pmatrix} I_k \\ B \end{pmatrix} = 0.$$

Corollary. *Interpreting B as a matrix with integer entries $0, 1$, the lattices $L(C)$ and $L(C^\perp)$ have generator matrices (corrspondingly)*

$$G = \begin{pmatrix} I_k & 0 \\ B & 2I_{n-k} \end{pmatrix}, \qquad H = \begin{pmatrix} -B^t & 2I_k \\ I_{n-k} & 0 \end{pmatrix}.$$

Let also denote the generator matrices for $\Lambda(C)$ and $\Lambda(C^\perp)$ correspondingly by

$$G_1 = \frac{1}{\sqrt{2}} G, \quad H_1 = \frac{1}{\sqrt{2}} H.$$

To show $\Lambda(C^\perp) = \Lambda(C)^\perp$ we first check $\Lambda(C^\perp) \subseteq \Lambda(C)^\perp$. This amounts to ${}^tH_1 G_1 = \frac{1}{2}{}^tHG$ being an integral matrix. We calculate

$$\frac{1}{2}{}^tHG = \frac{1}{2}\begin{pmatrix} -B & I_{n-k} \\ 2I_k & 0 \end{pmatrix}\begin{pmatrix} I_k & 0 \\ B & 2I_{n-k} \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 0 & 2I_{n-k} \\ 2I_k & 0 \end{pmatrix}.$$

But as we were so successful, namely $\frac{1}{2}{}^tHG = \left(\begin{smallmatrix} 0 & I_{n-k} \\ I_k & 0 \end{smallmatrix}\right)$, this calculation implies that $\Lambda(C^\perp) = \Lambda(C)^\perp$. This is because the last equality means that after permuting the columns of ${}^tH_1$ it becomes the inverse of G_1. By Exercise 19.4.2, it follows that the span of the columns of the permuted ${}^tH_1$, which is the span of the columns of ${}^tH_1$, is the dual lattice.

Consider now part (4). The self-duality of C implies $n = 2k$ and

$$I_k + {}^tBB = (I_k \ \ {}^tB)\begin{pmatrix} I_k \\ B \end{pmatrix} \equiv 0 \pmod 2.$$

Part (3) implies that $\Lambda(C)$ is self-dual if and only if C is self-dual, but we may also see that as follows: $\Lambda(C)$ is self-dual of type II if and only if $\mathrm{Span}_{\mathbb{Z}}(G_1) = \mathrm{Span}_{\mathbb{Z}}({}^tG_1^{-1})$ and the diagonal entries of ${}^tG_1G_1$ are even. The condition on the span holds if and only if $G_1 = {}^tG_1^{-1}N$ for some $N \in \mathrm{GL}_n(\mathbb{Z})$, which is equivalent to ${}^tG_1G_1 \in \mathrm{GL}_n(\mathbb{Z})$. A calculation gives

$${}^tG_1G_1 = \begin{pmatrix} \frac{1}{2}(I_k + {}^tBB) & {}^tB \\ B & 2I_{n-k} \end{pmatrix}.$$

Because $I_k + {}^tBB \equiv 0 \pmod 2$, ${}^tG_1G_1$ is an integral matrix. Its determinant is $\det(G_1)^2 = \frac{1}{2^n}\det(G)^2 = 1$, and so it belongs to $\mathrm{GL}_n(\mathbb{Z})$. (And, conversely, ${}^tG_1G_1$ being an integral matrix implies that $I_k + {}^tBB \equiv 0 \pmod 2$ and that C is self-dual.) Now, the fact that C is of type II means that the weight of every vector is divisible by 4. In particular, the diagonal entries of $I_k + {}^tBB = (I_k \ \ {}^tB)\begin{pmatrix} I_k \\ B \end{pmatrix}$ which express the length of the generators of C, are divisible by 4. It follows that the diagonal entries of ${}^tG_1G_1$ are even. This concludes the proof.

□

Exercise 25.1.6. Prove that the lattice E_8 is a unimodular lattice. Namely, E_8 is self-dual and $\mathrm{covol}(E_8) = 1$. Prove that the same is true for $\mathbb{Z}^8$. Prove that the kissing number of E_8 is 240, while for $\mathbb{Z}^8$ it is 16. This again illustrate how dramatically better the E_8 -packing is in comparison to the square packing provided by $\mathbb{Z}^8$.

26. Even unimodular lattices

In this section we discuss some striking results about unimodular lattices. Some are rather hard, for example Niemeier's Theorem, or the Siegel-Minkowski Theorem, while others, for example Witt's theorem, are rather accessible.[41]

26.1. **Some remarkable theorems concerning even unimodular lattices.** A lattice $\mathscr{L}$ is called **unimodular** if it is self-dual; equivalently, if it is integral and of covolume 1. We can obtain examples of such lattices from self-dual codes via Construction A. Such a lattice is called even, or type II, if $\|x\|^2 \in 2\mathbb{Z}$ for all $x \in \mathscr{L}$. If B is a Gram matrix for $\mathscr{L}$, this is the case if and only if B has even diagonal entries. An example of an even unimodular lattice is the E_8 lattice. If $n = 8n$ then the orthogonal sum of E_8 with itself n times, $E_8 \oplus \cdots \oplus E_8$, shows that there is an even unimodular lattice of dimension $8n$ for every $n \geq 1$. Remarkably, the converse holds.

Theorem 26.1.1 (E. Hecke)**.** *There is an even unimodular lattice of dimension n if and only if $8|n$.*

Theorem 26.1.2 (L. J. Mordell)**.** *Up to isometries, E_8 is the unique even unimodular lattice in $\mathbb{R}^8$.*

Theorem 26.1.3 (E. Witt)**.** *Up to isometries $E_8 \oplus E_8$ and D_{16}^+ are the unique even unimodular lattices in $\mathbb{R}^{16}$.*

Remark 26.1.4. It's an interesting point that although $E_8 \oplus E_8$ and D_{16}^+ are not isometric, they have the same theta function,

$$1 + 480 \sum_{m=1}^{\infty} \sigma_7(m) q^{2m},$$

where for a positive integer r we let $\sigma_r(m) = \sum_{d|m} d^r$ (the sum is over positive divisors of m, including 1 and m); namely, for every integer m they have exactly the same number of vectors of length m. It follows that if one considers $\mathbb{R}^{16}/E_8 \oplus E_8$ and $\mathbb{R}^{16}/D_{16}^+$ the spectrum of the Laplacian operator on these two manifolds is the same. They have the same harmonics, the same "sound". This example, provide by J. W. Milnor, was the first example providing a negative answer to the question "can one hear the shape of the drum?".

Remark 26.1.5. The lattices D_n^+ are an interesting example. We explain what they are in a series of exercises. Recall that

$$D_n = \{(x_1, \ldots, x_n) \in \mathbb{Z}^n : \sum x_i \equiv 0 \pmod 2\}.$$

Let $[½] = (½, ½, \ldots, ½) \in \mathbb{R}^n$. Let

$$D_n^+ = D_n \coprod ([½] + D_n).$$

Exercise 26.1.6.
(1) Prove that D_n^+ is a lattice if and only if n is even.
(2) Prove that D_n^+ is an integral lattice if and only if $4|n$.
(3) Prove that D_n^+ is even if and only if $8|n$.
(4) Prove that $\mathrm{covol}(D_n^+) = 1$.
(5) For n even, prove that $\mu_1(D_n) = \mu_1(D_n^+)$ and calculate $\delta(D_n^+)$.

It follows from Mordell's theorem that $D_8^+ \cong E_8$ (they are not actually equal). Both have the theta function

$$1 + 240 \sum_{m=1}^{\infty} \sigma_3(m) q^{2m}.$$

[41] The proof of Witt's theorem can be found in the book J.-P. Serre, *A Course in Arithmetic*, GTM 7.

Exercise 26.1.7. Find the vectors x in $E_8 \oplus E_8$ and D_{16}^+ such that $\|x\|^2 = 2$ (there should be 480 of them). Prove that in $E_8 \oplus E_8$ they generate the lattice while in D_{16}^+ they do not. Conclude that

$$E_8 \oplus E_8 \ncong D_{16}^+.$$

Theorem 26.1.8 (H.-V. Niemeier)**.** *Up to isometries, there are* 24 *even unimodular lattices in* $\mathbb{R}^{24}$*.*

Exercise 26.1.9. Prove that $E_8^3, E_8 \oplus D_{16}^+, D_{24}^+, \Lambda(\mathscr{G}_{24}), \Lambda_{24}$ are examples. (You are not required to prove they are mutually non-isomorphic although this is true). Here Λ_{24} is the Leech lattice already defined in § 25.1.1.

In higher dimensions exploring lattices is a bit like exploring star systems in the universe. For example, it is known that there are more than $80,000,000$ even unimodular lattices in dimension 32. Finding an interesting one is a bit like finding life on another planet.

Theorem 26.1.10 (Minkowski-Siegel)**.** *Let*

$$M_n = \sum_{\mathscr{L}} \frac{1}{\sharp \mathrm{Aut}(\mathscr{L})},$$

the summation extending over all even unimodular lattices in dimension $n = 8k$, $k > 0$. Then,

$$M_n = \frac{B_{2k}}{8k} \prod_{i=1}^{4k-1} \frac{B_j}{4j},$$

where B_s are the Bernoulli numbers.[42]

For example, if one calculates the isometry group, also called its automorphism group, of E_8, namely, all the isometries of $\mathbb{R}^8$ that map E_8 isomorphically onto itself, and finds that its order is 696729600, one can conclude that E_8 is the unique even modular lattice of dimension 8, up to isometry. Or, conversely, if one allows Hecke's theorem that E_8 is the unique even modular lattice of dimension 8, up to isometry, then one can calculate the size of its isometry group.

As a final remark, the automorphism groups of lattices turn out to be very interesting objects. Conway discovered three new sporadic groups, and reconstructed already known sporadic groups, from automorphism groups of lattices.

26.2. **The Leech lattice.** It is rather unfortunate that there is no really simple way to introduce the Leech lattice. Although many different constructions are known, for example some use copies of the E_8 lattice, some use the Octonions, some use the Hamilton quaternions over $\mathbb{Q}[\sqrt{5}]$, none is straightforward. Of course, as the Leech lattice is a lattice in $\mathbb{R}^{24}$ one could simply list a basis for the lattice consisting of 24 vectors, but this is hardly illuminating. The construction we present at least starts very conceptually, by applying Construction A to the extended Golay code $\mathscr{G}_{24}$, but then a rather unmotivated tweaking is required.

Recall the $(24, 8, 8)$ extended Golay code $\mathscr{G}_{24} = \mathscr{G}_{23}^e$. It is a doubly even, self dual code. Let

$$\mathbb{L} = L(\mathscr{G}_{24}) = \{a \in \mathbb{Z}^{24} : a \mod 2 \in \mathscr{G}_{24}\}.$$

[42]The theorem can be found in Conway & Sloane, Chapter 16, but our convention for the Bernoulli numbers is as in Serre, *A course in Arithmetic*. For example: $B_1 = \frac{1}{6}, B_2 = \frac{1}{30}, B_3 = \frac{1}{42}, B_4 = \frac{1}{30}, B_5 = \frac{5}{66}, B_6 = \frac{691}{2730}, B_7 = \frac{7}{6}$

As the co-volume of $\Lambda(C)$ is 1, the co-volume of $\mathbb{L}$ is 2^{12}, and as $\Lambda(C)$ is self-dual, the dual of $\mathbb{L} = \sqrt{2} \cdot \Lambda(C)$ is $\mathbb{L}^\perp = \frac{1}{2}\mathbb{L}$. Define a new lattice

$$\mathbb{B} = \{v \in \mathbb{L} : \sum_{i=1}^{24} v_i \equiv 0 \pmod 4\}$$

Then, $\mathbb{B}$ is a sublattice of $\mathbb{L}$. The map

$$\mathbb{L} \to \mathbb{Z}/4\mathbb{Z}, \quad v \mapsto \sum v_i \pmod 4$$

has image $\{0, 2\}$ and $\mathbb{B}$ is its kernel. It follows that $[\mathbb{L} : \mathbb{B}] = 2$.

Let

$$t = \frac{1}{2}(-3, 1, 1, \ldots, 1) \in \frac{1}{2}\mathbb{Z}^{24}.$$

Define

$$\tilde{\mathbb{L}} = \mathbb{B} \coprod (t + \mathbb{B}).$$

Proposition 26.2.1. *$\tilde{\mathbb{L}}$ has the following properties.*

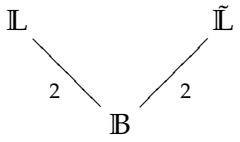
(1) $\tilde{\mathbb{L}}$ is a lattice and $[\tilde{\mathbb{L}} : \mathbb{B}] = 2$.

(2) $\frac{1}{\sqrt{2}}\tilde{\mathbb{L}}$ is an even self-dual lattice. The **Leech lattice** *Λ_{24} is defined as*

$$\Lambda_{24} = \frac{1}{\sqrt{2}}\tilde{\mathbb{L}}.$$

It is thus an even unimodular lattice.

Proof. We first note that $v = 2t$ satisfies $\sum_i v_i = -3 + 23 \cdot 1 \equiv 0 \pmod 4$. To check $v \in \mathbb{L}$ we note that mod 2, $v = (1, 1, \ldots, 1)$ which is in $(\mathscr{G}^{24})^\perp$ because very code word in $\mathscr{G}^{24}$ is even. Since $\mathscr{G}^{24}$ is self-dual, $(1, 1, \ldots, 1) \in \mathscr{G}^{24}$. Therefore, $v \in \mathbb{B}$. As $-t = t - v$, it follows that $\tilde{\mathbb{L}}$ is an abelian group and so a lattice; furthermore, $[\tilde{\mathbb{L}} : \mathbb{B}] = 2$. We have the following situation:



In particular, $\mathrm{covol}(\tilde{\mathbb{L}}) = \mathrm{covol}(\mathbb{L}) = 2^{12}$.

We now check that Λ_{24} is even and integral. We need to show that for all $x, y \in \Lambda_{24}$ we have $x \cdot y \in \mathbb{Z}$ and $\|x\|^2 = x \cdot x \in 2\mathbb{Z}$. This is equivalent to showing that for all $x, y \in \tilde{\mathbb{L}}$ we have $x \cdot y \in 2\mathbb{Z}$ and $\|x\|^2 = x \cdot x \in 4\mathbb{Z}$.

Let us think about $\mathbb{L}$ as

$$\mathbb{L} = \cup_{c \in \mathscr{G}_{24}} c + 2\mathbb{Z}^{24},$$

where c is taken to be a vector with coordinates 0 and 1. Every code word in $\mathscr{G}_{24}$ has weight divisible by 8. Therefore,

$$\mathbb{B} = \{c + u : c \in \mathscr{G}_{24}, u \in 2\mathbb{Z}^{24}, \sum_i u_i \equiv 0 \pmod 4\}.$$

From this follows

$$\tilde{\mathbb{L}} = \{\epsilon t + c + u : c \in \mathscr{G}_{24}, u \in 2\mathbb{Z}^{24}, \sum_i u_i \equiv 0 \pmod 4, \epsilon \in \{0, 1\}\}.$$

It is useful to note that $t = \frac{1}{2}(1, 1, \ldots, 1) - (2, 0, \ldots, 0)$. Using that $\mathscr{G}_{24}$ is doubly even, we find that

$$t \cdot t = 8, \qquad t \cdot c \equiv 0 \pmod 2, \quad t \cdot u \equiv 0 \pmod 2,$$
$$c_1 \cdot c_2 \equiv 0 \pmod 2, \quad c \cdot u \equiv 0 \pmod 2, \quad u \cdot u' \equiv 0 \pmod 4.$$

Thus, for all $x, y \in \tilde{\mathbb{L}}$ we have $x \cdot y \in 2\mathbb{Z}$ and

$$x \cdot x = (\epsilon t + c + u) \cdot (\epsilon t + c + u) = \epsilon^2 t \cdot t + c \cdot c + u \cdot u + 2\epsilon t \cdot c + 2\epsilon t \cdot u + 2c \cdot u \equiv 0 \pmod 4,$$

the critical point being that $c \cdot c \equiv 0 \pmod 4$ because $\mathscr{G}_{24}$ is doubly even.

We have thus far shown that the Leech lattice Λ_{24} is an even integral lattice. As $\mathrm{covol}(\Lambda_{24}) = \frac{1}{2^{12}}\mathrm{covol}(\tilde{\mathbb{L}}) = 1$, we conclude that Λ_{24} is unimodular. □

Let $\mathscr{L}$ be a unimodular lattice in $\mathbb{R}^n$. The packing density of $\mathscr{L}$ is then equal to

$$\Delta(\mathscr{L}) = \frac{\mu_1(\mathscr{L})^n \omega_n}{2^n \mathrm{covol}(\mathscr{L})} = \mu_1(\mathscr{L})^n \times \frac{\omega_n}{2^n}.$$

Thus, all unimodular lattices that have some given $\mu_1(\mathscr{L})$ provide the same packing density. From Niemeier's result, we know that there are 24 even unimodular lattices in $\mathbb{R}^{24}$, among which are $\Lambda(\mathscr{G}_{24})$ and the Leech lattice Λ_{24}. The minimum of $\|x\|^2$ for $x \in \Lambda(\mathscr{G}_{24})$ is 2 (take a code word c of $\mathscr{G}_{24}$ of weight 8 and view $\frac{1}{\sqrt{2}}c$ as an element of $\Lambda(\mathscr{G}_{24})$; it has norm squared equal to 2 and so $\mu_1 = \sqrt{2}$. Thus $\Lambda(\mathscr{G}_{24})$ and any other even unimodular lattice with $\mu_1 = \sqrt{2}$ will have the same packing density. For an even unimodular lattice, μ_1 cannot be smaller. It turns out that among the even unimodular lattices in dimension 24 *there is only one*, the Leech lattice, that has larger μ_1. In fact, it will have $\mu_1 = 2$.

Theorem 26.2.2. *Let Λ_{24} be the Leech lattice. Then*

$$\mu_1(\Lambda_{24}) = 2, \quad \delta(\Lambda_{24}) = 1.$$

Proof. The statement about the centre density $\delta(\Lambda_{24})$ is a direct consequence of the statement about μ_1 and that Λ_{24} is unimodular. Thus, it will be enough to prove that $\mu_1(\tilde{\mathbb{L}}) = 2\sqrt{2}$, equivalently

$$\forall x \neq 0, x \in \tilde{\mathbb{L}}, \qquad \|x\|^2 = 8.$$

Take first an $0 \neq x \in \mathbb{B}$, $x = c + u$, in the notation of Proposition 26.2.1. Then

$$x = (c_1 + u_1, \cdots, c_{24} + u_{24}), \qquad |c_i + u_i| \geq |c_i|.$$

Therefore, since every non-zero code word in $\mathscr{G}_{24}$ has weight at least 8, we get for $c \neq 0$

$$\|x\|^2 \geq \|c\|^2 \geq 8,$$

and for $c = 0$, we have

$$\|x\|^2 = \|u\|^2 \equiv 0 \pmod 8,$$

because u has only even integers entries and, because $\sum u_i \equiv 0 \pmod 4$, there is an even number of them.

Now consider x of the form $t + c + u$. Then $2x = 2t + 2c + 2u$ and $2t = (1, 1, \ldots, 1) - (4, 0, \ldots, 0)$. As $2x \in \sqrt{8}\Lambda_{24}$ and Λ_{24} is an even integral lattice, if $\|x\|^2 < 32$ then it can only be 16. This implies that the number of non-zero entries of $2x$ is atmost 16. But, since every entry of $2c + 2u$ is even and every entry of t is odd, in fact all 24 entries of $2x$ are non-zero. □

The passage from $\mathbb{L}$ to $\tilde{\mathbb{L}}$ is a process one can try in general (and is studied in Conway & Sloane), and it is a fair question to ask if one cannot get a lattice with even larger μ_1. This is a good question, but the Roger's bound shows that the answer is "no" in dimension 24 (this argument doesn't require the optimality of the Leech lattice proven by Viazovska et al.).

.

APPENDICES

APPENDIX A. CHEAT SHEET FOR GAUSS AND JACOBI SUMS

$\chi : \mathbb{F}_q^\times \to \mathbb{C}^\times$ $\quad \psi(x) = \zeta_p^{\mathrm{Tr}_{\mathbb{F}_q/\mathbb{F}}(x)}$ $\quad \zeta_p = e^{2\pi i/p}$

$\mathfrak{g}_a(\chi) = \sum_{t\in\mathbb{F}} \chi(t)\psi(at)$ $\quad \mathfrak{g}(\chi) = \sum_{t\in\mathbb{F}} \chi(t)\psi(t)$ $\quad \mathfrak{g}_a(\chi) = \chi(a^{-1})\mathfrak{g}(\chi), a \neq 0, \chi \neq \epsilon$ $\quad$ and $\mathfrak{g}_0(\epsilon) = q$; else 0.

$|\mathfrak{g}(\chi)| = \sqrt{q},$ if $\chi \neq \epsilon$ $\quad \overline{\mathfrak{g}(\chi)} = \chi(-1)\mathfrak{g}(\bar{\chi})$ $\quad \mathfrak{g}(\chi)\mathfrak{g}(\bar{\chi}) = \chi(-1)q$

$J(\chi_1,\dots,\chi_\ell) = \sum_{t_1+\cdots+t_\ell=1} \chi_1(t_1)\cdots\chi_\ell(t_\ell)$ $\quad J_0(\chi_1,\dots,\chi_\ell) = \sum_{t_1+\cdots+t_\ell=0} \chi_1(t_1)\cdots\chi_\ell(t_\ell)$

$J(\epsilon,\dots,\epsilon) = q^{\ell-1}$ $\quad J_0(\epsilon,\dots,\epsilon) = q^{\ell-1}$

$J_0(\chi_1,\dots,\chi_\ell) = 0$ $\quad J(\chi_1,\dots,\chi_\ell) = 0$ $\quad$ if $\exists i, \neg\forall i, \chi_i = \epsilon$

$J_0(\chi_1,\dots,\chi_\ell) = 0$ $\quad$ if $\chi_\ell \neq \epsilon$ and $\chi_1\cdots\chi_\ell \neq \epsilon$

$J_0(\chi_1,\dots,\chi_\ell) = \chi_\ell(-1)(q-1)J(\chi_1,\dots,\chi_{\ell-1})$ $\quad$ if $\chi_\ell \neq \epsilon$ and $\chi_1\cdots\chi_\ell = \epsilon$ $\quad |J_0(\chi_1,\dots,\chi_\ell)| = (q-1)q^{\frac{\ell-2}{2}}$

$J(\chi_1,\dots,\chi_\ell) = \mathfrak{g}(\prod_{i=1}^\ell \chi_i)^{-1}\prod_{i=1}^\ell \mathfrak{g}(\chi_i)$ $\quad \forall i\ \chi_i \neq \epsilon$ and $\prod_{i=1}^\ell \chi_i \neq \epsilon$ $\quad |J(\chi_1,\dots,\chi_\ell)| = q^{\frac{\ell-1}{2}}$

$J(\chi_1,\dots,\chi_\ell) = -\chi_\ell(-1)J(\chi_1,\dots,\chi_{\ell-1})$ $\quad \forall i\ \chi_i \neq \epsilon$ and $\prod_{i=1}^\ell \chi_i = \epsilon$ $\quad |J(\chi_1,\dots,\chi_\ell)| = q^{\frac{\ell-2}{2}}$

$J(\chi_1,\dots,\chi_{\ell-1}) = \frac{1}{q}\chi_\ell(-1)\prod_{i=1}^\ell \mathfrak{g}(\chi_i)$ $\quad \forall i\ \chi_i \neq \epsilon$ and $\prod_{i=1}^\ell \chi_i = \epsilon$

APPENDIX B. SOME USEFUL CONSTANTS

B.1. **π and e and some square roots.** The following values are truncated, not rounded.

$$\pi = 3.14159265358979323846$$

$$e = 2.71828182845904523536$$

$$\sqrt{2} = 1.4142135623, \quad \sqrt{3} = 1.7320508075, \quad \sqrt{5} = 2.2360679774$$

B.2. **Volumes of balls.** The following table provides the volume ω_n of the unit ball in $\mathbb{R}^n$.

n	1	2	3	4	5	6	7	8	9	10	11	12	16	24
$\mathrm{vol}(B_n[0,1])$	2	π	$\frac{4\pi}{3}$	$\frac{\pi^2}{2}$	$\frac{8\pi^2}{15}$	$\frac{\pi^3}{6}$	$\frac{16\pi^3}{105}$	$\frac{\pi^4}{24}$	$\frac{32\pi^4}{945}$	$\frac{\pi^5}{120}$	$\frac{64\pi^5}{10395}$	$\frac{\pi^6}{720}$	$\frac{\pi^8}{40320}$	$\frac{\pi^{12}}{479001600}$

General formulas are given by:

$$\omega_n = \begin{cases} \frac{\pi^k}{k!}, & n = 2k \text{ even;} \\ \frac{2^{2k+1}k!\pi^k}{(2k+1)!} & n = 2k+1 \text{ odd.} \end{cases}$$

and the recursion formula

$$\omega_1 = 2, \ \omega_2 = \pi, \quad \omega_n = \omega_{n-2}\frac{2\pi}{n}.$$

B.3. **Bernoulli numbers.** The following tables gives the values of the Bernoulli numbers defined by the identity

$$\frac{x}{e^x - 1} = 1 - \frac{x}{2} + \sum_{k=1}^{\infty} (-1)^{k+1} B_k \frac{x^{2k}}{(2k)!}.$$

There are other normalizations for Bernoulli numbers in the literature. Ours follows Serre in A course in arithmetic, but note that Conway & Sloane use different conventions in *Sphere packings, lattices and groups*.

n	1	2	3	4	5	6	7	8	9	10	11	12
B_n	$\frac{1}{6}$	$\frac{1}{30}$	$\frac{1}{42}$	$\frac{1}{30}$	$\frac{5}{66}$	$\frac{691}{2730}$	$\frac{7}{6}$	$\frac{3617}{510}$	$\frac{43867}{798}$	$\frac{174{,}611}{330}$	$\frac{854513}{138}$	$\frac{236364091}{2730}$

EXERCISES

(1) What is the continued fraction $[1,2,3,1,2,3,1,2,3,\dots]$?

(2) Let a be a positive integer. What is the continued fraction expansion of the positive root of $x^2 - ax - 1$?

(3) Let a be a positive integer. What is the continued fraction expansion of the positive root of $x^2 + ax - 1$?

(4) Fill in the following table for $e = \exp(1)$. You may use a calculator, or a computer software. (For $1/q_n^2$ write an approximate decimal expansion.)

n	$[a_0, a_1, \dots, a_n]$	p_n/q_n	$e - p_n/q_n$	$1/q_n^2$	optimal?
0	[2]	2/1	0.7182818...	1.0	no
1	[2, 1]	3/1	-0.2817181...	1.0	yes
2					
3					
4					
5					

(5) Using GP-PARI (see https://pari.math.u-bordeaux.fr), or any other mathematical software, or even an online calculator (but make sure it's precise enough otherwise you may be lead to a wrong conjecture), find the continued fractions expansions of

$$\frac{e-1}{2}, \quad \frac{e^{1/2}-1}{2}, \quad \frac{e^{1/3}-1}{2}, \quad \frac{e^{1/4}-1}{2}, \dots$$

Formulate a conjecture.

(6) Let $[a_0, a_1, a_2, \dots]$ be a continued fraction, where, as usual $a_0 \in \mathbb{Z}, a_i \in \mathbb{N}^+, i = 1, 2, 3, \dots$. Prove that

$$q_n \geq 2^{\frac{n-1}{2}}.$$

(7) Prove that if $a_0, b_0 \in \mathbb{Z}$, $a_i, b_i \in \mathbb{N}^+$ for $i \geq 1$, we cannot have

$$[a_0, \dots, a_n] = [b_0, b_1, b_2, \dots].$$

(8) Prove that every rational number θ has a finite continued fraction expansion

$$\theta = [a_0, a_1, \dots, a_N] \quad (a_0 \in \mathbb{Z}, a_i \in \mathbb{N}^+, i = 1, \dots, N).$$

Moreover, prove that this expansion is *unique*, up to

$$[a_0, a_1, \dots, a_N] = [a_0, a_1, \dots, a_{N-1}, 1],$$

if $a_N > 1$.

(9) ★ Use the arguments appearing in Theorem 2.3.1 in the notes to prove
Theorem 12. Let θ be an irrational real number. Then, for all $n \geq 0$ we have

$$\left|\theta - \frac{p_n}{q_n}\right| > \frac{1}{q_n(q_{n+1} + q_n)}.$$

(10) Prove that the measure of $[0,1] \setminus \mathbb{Q}$ is equal to 1. Note that this set contains no interval of positive length.

(11) Let $0 \leq \alpha \leq 1$. Find a set S contained in $[0,1]$ that has measure α, contains no interval of positive length, and is dense in $[0,1]$.

(12) ★ Prove that $\bar{\mathbb{Q}}$ is a field as follows:
   (a) In general, if $F \subseteq L$ are commutative rings and $\alpha \in L$ let
   $$F[\alpha] = \{\sum_{i=0}^{n} a_i \alpha^i : a_i \in \mathbb{F}\}.$$
   Namely, the set of all finite polynomial expression in α with coefficients from F. Prove that $F[\alpha]$ is a ring. If F is a field prove that it is also a vector space over F.
   (b) Prove that $\alpha \in \mathbb{C}$ is algebraic over $\mathbb{Q}$ if and only if $\dim_{\mathbb{Q}}(\mathbb{Q}[\alpha]) < \infty$. If this is the case, prove that $\mathbb{Q}[\alpha]$ is a field and, in fact, $\mathbb{Q}[\alpha] \cong \mathbb{Q}[x]/(f(x))$, where $f(x)$ is the minimal polynomial of α.
   (c) Let $\alpha, \beta \in \mathbb{C}$ be algebraic over $\mathbb{Q}$. Prove that $\dim_{\mathbb{Q}}(\mathbb{Q}[\alpha,\beta]) < \infty$, where $\mathbb{Q}[\alpha,\beta] = (\mathbb{Q}[\alpha])[\beta]$.
   (d) Let $\alpha, \beta \in \mathbb{C}$ be algebraic over $\mathbb{Q}$. Prove that $-\alpha$, $\frac{1}{\alpha}$ (for $\alpha \neq 0$), $\alpha + \beta$ and $\alpha\beta$ all belong to $\mathbb{Q}[\alpha,\beta]$. Conclude that they are algebraic too.

(13) Prove that $\log(2), \log(3), \log(2) + \log(3), \log(2)/\log(3)$ are transcendental numbers.

(14) ★ Use the expansion
$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \dots$$
to prove that e is not a rational number (which is much easier than proving it's transcendental!).

(15) Show that there are no inverse implications in Lemma 4.1.1.

(16) Prove that every real irrational number θ has infinitely many BAF without using Dirichlet's theorem. (Or continued fractions...)

(17) Prove that every real irrational number θ has infinitely many BAS by using Dirichlet's theorem.

(18) Analyze the proof of Liouville's theorem and find a constant C as in the theorem for $\sqrt{2}$, $\frac{1+\sqrt{5}}{2}$, $\sqrt[3]{5} \in \mathbb{R}$.

(19) Let $\theta = \sum_{n=1}^{\infty} \frac{1}{10^{n!}}$. Prove that θ is transcendental.

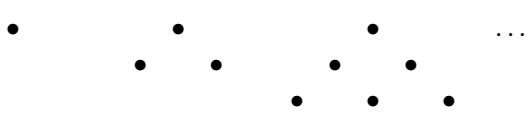
(20) ★ Construct a set T of real transcendental numbers such that $\left|T\right| > \aleph_0$ and $\mu(T) = 0$.

(21) Find positive solutions for the following equations:
   (a) $x^2 - 39y^2 = 1$.
   (b) $x^2 - 41y^2 = 1$.

(22) Prove that there are infinitely many solutions to the equation
$$x^2 - 39y^2 = -3.$$
(Hint: given a solution (a,b) to $x^2 - 39y^2 = -3$ and a solution (c,d) to $x^2 - 39y^2 = 1$, show that one can generate a new solution to $x^2 - 39y^2 = -3$ by using the product $(a + b\sqrt{39})(c + d\sqrt{39})$.)

(23) Find a positive solution to the equation $x^2 - 41y^2 = 5$.

(24) ★ **Triangular numbers** are the integers $1, 3, 6, \dots, \frac{n(n+1)}{2}, \dots$.



Show that there are infinitely many triangular numbers that are squares and find 3 of them besides $0, 1$.

(25) ★ Find five pairs of integers (n, N), $1 \leq n \leq N$, such that
$$1 + 2 + \dots + (n-1) = (n+1) + (n+2) + \dots + N.$$

(26) Let (a,b) be a solution to Pell's equation $x^2 - dy^2 = 1$. Show that for any n, if we define A_n, B_n as follows
$$A_n + B_n\sqrt{d} = (a + b\sqrt{d})^n,$$
then A_n, B_n are also solutions to the same equation. Use this to show that if a Pell equation $x^2 - dy^2 = N$ has a solution then it has infinitely many solutions.

(27) ★ Show that there are infinitely many solutions (a,b) to $x^2 - 10y^2 = 1$ such that $7|a$.

(28) Let d be an integer that is not a square.
   (a) The equation $x^2 - dy^2 = -1$ doesn't always have integral solutions: prove that if $d \equiv 0, -1 \pmod 4$ there are no integral solutions. However, prove that if a solution exists then it is a convergent to $\sqrt{d}$.
   (b) More generally, if $-\sqrt{d} < N < 0$ prove that every positive solution to the equation $x^2 - dy^2 = N$ is a convergent to $\sqrt{d}$.

(29) Prove that $[a, \overline{b, c,}\,] = \sqrt{n}$ if and only if $a > 0, c = 2a, b|c$, in which case $n = a^2 + c/b$.

(30) Find $\mu(S)$ and $\mu(T)$ where
$$S = \{[0, a_1, a_2, \dots] : a_1 = 2, a_2 = 3\}, \quad T = \{[0, a_1, a_2, \dots] : a_1 = a_2\}.$$
For T, the answer should be expressed as an infinite sum to which you should provide non-trivial (i.e. different that 0 or 1) lower and upper bounds.

(31) Prove that
$$\rho := \frac{\mu\left(E\left(\begin{smallmatrix}1 & 2 & \dots & n & n+1\\ k_1 & k_2 & \dots & k_n & s\end{smallmatrix}\right)\right)}{\mu\left(E\left(\begin{smallmatrix}1 & 2 & \dots & n\\ k_1 & k_2 & \dots & k_n\end{smallmatrix}\right)\right)} = \frac{1}{s^2}\cdot\frac{1+\frac{q_{n-1}}{q_n}}{(1+\frac{q_{n-1}}{sq_n})(1+\frac{1}{s}+\frac{q_{n-1}}{sq_n})};$$
it satisfies,
$$\frac{1}{3s^2} < \rho < \frac{2}{s^2},$$
independently of $k_1, \dots, k_n$ (!)

(32) Prove that $\mu(E\left(\begin{smallmatrix}1\\2\end{smallmatrix}\right)) = \frac{1}{6} = 0.1666\dots$, while
$$\mu(E\left(\begin{smallmatrix}2\\2\end{smallmatrix}\right)) = \sum_{k=1}^{\infty}\frac{1}{(2k+1)(3k+1)}.$$
The value of this series is numerically close to 0.1685. How well can you approximate this sum?

(33) ★ Let $T\colon [0,1] \to [0,1]$ be the transformation given on continued fractions by $T([0, a_1, a_2, a_3, \dots] = [0, a_2, a_3, \dots]$. Prove that for any $0 \le \beta \le 1$,
$$\nu(T^{-1}(0,\beta)) = \nu((0,\beta)).$$

(34) Let $f(x) = \begin{cases}1 & a_1(x) = k\\ 0 & \text{else.}\end{cases}$ Using the ergodic theorem, deduce that for almost all $x \in (0,1)$ the frequency of k in the partial quotients of x, namely in the sequence $\{a_i(x)\}_{i=1}^{\infty}$, is
$$\frac{1}{\log(2)}\log\left(\frac{(k+1)^2}{k(k+2)}\right).$$

(35) Let $f(x) = a_1(x)$ and deduce that with probability 1,
$$\lim_{n\to\infty}\frac{a_1(x) + \dots + a_n(x)}{n} = \infty.$$

(36) ★ What result can we deduce from the Ergodic Theorem if we let $f(x) = 1$ if $a_1(x)$ is prime and $f(x) = 0$ otherwise?

(37) ★ Give a proof based on the Ergodic Theorem for Theorem 7.2.1.

(38) Prove that for every s, $\mathcal{H}^s(\cup_{i=1}^{\infty} F_i) \leq \sum_{i=1}^{\infty} \mathcal{H}^s(F_i)$.

(39) Prove parts (1) - (4) of Theorem 8.1.4. You may use Exercise 38.

(40) Let $f\colon \mathbb{R}^n \to \mathbb{R}^n$ be a function such that for some $\alpha > 0$

$$\|f(x) - f(y)\| = \alpha \|x - y\|.$$

Then,

$$\mathcal{H}^s(f(F)) = \alpha^s \mathcal{H}^s(F).$$

(41) Let A be the set of all numbers in $[0,1]$ whose base 5 expansion only contains the digits $0, 2$ and 4. Let $d = \dim_H(A)$ and assume that $0 < \mathcal{H}^d(A) < \infty$. Calculate $\dim_H(A)$.

(42) ★ Let $N \geq 3$ be an odd integer. Let A_N be the set of all numbers in $[0,1]$ whose base N expansion only contains the digits $0, 2, \ldots N-1$. Let $d = \dim_H(A_N)$ and assume again that $0 < \mathcal{H}^d(A_N) < \infty$. Calculate $\lim_{N \to \infty} \dim_H(A_N)$.

(43) Let $F \subseteq \mathbb{R}^n$ be a subset such that $\dim_H(F) < 1$. Prove that F is totally disconnected. Here is a suggestion. Suppose $x \neq y$ are points in F:

- Define
$$f\colon \mathbb{R}^n \to \mathbb{R}, \quad f(s) = \|x - s\|.$$
Prove that this function has the property
$$|f(s) - f(t)| \leq \|s - t\|.$$
- Prove that $\dim_H(f(F)) \leq \dim_H(F)$.
- Let $Z = \mathbb{R} \setminus f(F)$. Prove that Z is dense in $\mathbb{R}$.
- Prove that there is a $z \in Z$ lying between $f(x)$ and $f(y)$ and so the sets $(-\infty, z), (z, \infty)$ separate $f(x)$ and $f(y)$.
- Complete the proof.

(44) Let D be a compact set and let $\mathcal{S}$ be the collection of non-empty compact subsets of D. We will make S into a metric space; the metric is known as the **Hausdorff metric**. Let $\delta \geq 0$. Define the **δ-neighbourhood** of a set $A \in \mathcal{S}$, denoted A_δ, to be the set

$$A_\delta = \{x \in D : \exists a \in A, |x - a| \leq \delta\}.$$

Using that for a fixed $x \in D$, $\inf_{a \in A}\{\|x - a\|\}$ is achieved for some $a_x \in A$, by compactness of A, it is not hard to prove that A_δ is a closed subset of D, hence in $\mathcal{S}$ itself.

Now, given $A, B \in \mathcal{S}$, define

$$d(A, B) = \inf\{\delta : A \subseteq B_\delta \text{ and } B \subseteq A_\delta\}.$$

(45) Prove that for $A_i, B_j \in \mathcal{S}$,

$$d(\cup_{i=1}^{m} A_i, \cup_{j=1}^{m} B_j) \leq \max_{1 \leq i \leq m} d(A_i, B_i).$$

(46) Prove that the attractor E of an IFS is unique in the following sense. If $F \subseteq D$ is a compact non-empty set of D such that $S^1(F) = F$ then $F = E$. (Consider $d(E, F)$ and apply the previous exercises for $A_i = s_i(E), B_i = s_i(F)$.)

(47) Prove that the dimension of the Sierpinski cube is $\log(20)/\log(3)$. (See Figure 12. Picture from Wikipedia commons.)

(48) Consider a set $\mathscr{C}_\alpha$ which is very similar to the Cantor set. At each step we remove an interval of length 2α which is centrally located. See Figure 13. Thus, the case of the Cantor set itself is when $\alpha = 1/3$. Calculate the cardinality, the measure and the dimension of $\mathscr{C}_\alpha$.

(49) ★ Use Ramharter's Theorem to prove that $\dim_H(S) \geq 0.9$, say, where S is the set of all real numbers for which the partial quotients of their continued fractions take only finitely many values. The theorem is not powerful enough to imply that $\dim_H(S) = 1$, but it can be used for a great many examples of sets defined by conditions on continued fractions.

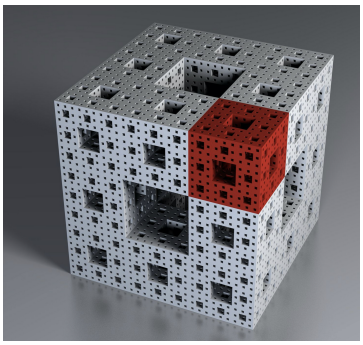


FIGURE 12. The Sierpinski cube.

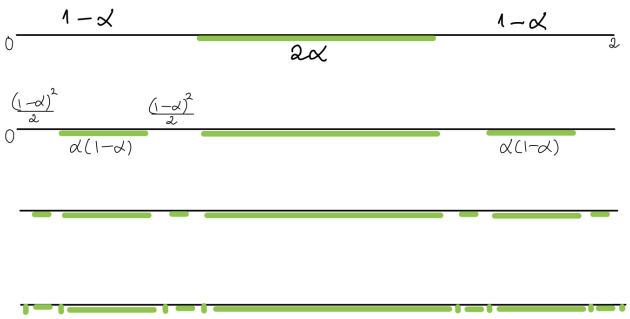


FIGURE 13. Generalized Cantor set $\mathscr{C}_\alpha$.

(50) ★ Let $\mathscr{C}$ be the Cantor set then
$$\mathscr{C}+\mathscr{C}=\{x+y : x,y\in\mathscr{C}\}=[0,2].$$

(51) Let $\mathbb{L}/\mathbb{F}$ be a finite extension of finite fields. Prove that the maps trace $\mathrm{Tr}_{\mathbb{L}/\mathbb{F}}$ and norm $\mathrm{Nm}_{\mathbb{L}/\mathbb{F}}$ are surjective maps onto $\mathbb{F}$ and $\mathbb{F}^\times$, respectively.

(52) Let $P=P_{m,n}$ be the subgroup of GL_n such that $P(\mathbb{L})$ consists of matrices M of the form $M=\left(\begin{smallmatrix}A & B\\ 0 & D\end{smallmatrix}\right)$, $A\in\mathrm{GL}_m(\mathbb{L}), B\in M_{m,n-m}(\mathbb{L}), D\in\mathrm{GL}_{n-m}(\mathbb{L})$. Show that for a field $\mathbb{L}$ there is a natural bijection
$$\mathrm{GL}_n(\mathbb{L})/P(\mathbb{L})\leftrightarrow G_{m,n}(\mathbb{L}).$$
This bijection is associating to a right coset $M\cdot P(\mathbb{L})$ the m-dimensional subspace of $\mathbb{A}^n$ spanned by the first m columns of M.

(53) Let $\mathbb{L}$ be a finite field with q elements. Prove that
$$\sharp\mathrm{GL}_n(\mathbb{L})=\prod_{i=1}^{n}(q^n-q^{n-i})=:c(n),$$
$$\sharp G_{m,n}(\mathbb{L})=\frac{c(n)}{c(m)c(n-m)q^{m(n-m)}}.$$
Verify the formula for the case of the projective space.

(54) Prove that the Grassmann variety $G_{2,4}$, considered as a variety over $\mathbb{F}_p$, has the following zeta function:
$$\zeta(T)=\frac{1}{(1-T)(1-pT)(1-p^2T)^2(1-p^3T)(1-p^4T)}.$$

(55) Calculate the zeta function of the projective surface $x_0x_1-x_2x_3=0$ over $\mathbb{F}_p$, in $\mathbb{P}^3$.

(56) Find the number of projective points of the elliptic curve $y^2 = x^3 - 1$ over $\mathbb{F}_5$. Use it to calculate the cardinalities of $\mathbb{E}(\mathbb{F}_{5^2}), \mathbb{E}(\mathbb{F}_{5^3}), \mathbb{E}(\mathbb{F}_{5^4})$; write the zeta function of E as a ratio of explicit polynomials.

(57) In the context of the Sato-Tate conjecture, find the probability that

$$|E_p(\mathbb{F}_p) - (p+1)| \leq \sqrt{p}.$$

(58) Prove Lemma 13.1.4. Compare with the proof of Lemma 13.1.1.

(59) Let p be an odd prime and q a power of p. Prove the formula $N(x^2 = a) = 1 + \left(\frac{\mathrm{Nm}(a)}{p}\right)$ for $\mathbb{F}_q$ by proving that a is a square in $\mathbb{F}_q$ if and only if $\mathrm{Nm}(a)$ is a square in $\mathbb{F}_p$.

(60) Consider the case of $\mathbb{F}_p$ and let $\lambda(a) = \left(\frac{a}{p}\right)$ be the Legendre symbol. Suppose that $p \nmid a$. By considering two ways to evaluate the sum $\sum_{n=0}^{p-1} \left(1 + \left(\frac{n}{p}\right)\right) e^{2\pi i a n/p}$, prove that

$$\mathfrak{g}(\lambda) = \sum_{n=0}^{p-1} e^{\frac{2\pi i a n^2}{p}}.$$

(61) How many solutions do the following equations have?
(a) $x^2 + 120 \equiv 0 \pmod{257}$.
(b) $x^2 - x - 1 \equiv 0 \pmod{p}$, where $p > 5$ is a prime.

(62) Find a prime $p > 2$ such that $1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11$ are all squares modulo p. You may use a computer for some of the computations.

(63) Let F_n be the n-th Fermat number, $n \geq 1$.
(a) Prove that $F_n \equiv 5 \pmod{12}$.
(b) If F_n is prime, prove that

$$\left(\frac{3}{F_n}\right) = -1.$$

(64) Show that if $q \equiv 2 \pmod 3$ then

$$N(x^3 + y^3 = 1) = q.$$

(65) ★ Prove, using a suitable Jacobi sum, that if $p \equiv 1 \pmod 3$ then for suitable integers a, b

$$p = a^2 - ab + b^2.$$

(66) Let $p \equiv 1 \pmod 3$ be a prime. Let χ be an order 3 character in $\mathbb{X}_p$. Let $A = 2\mathrm{Re}J(\chi, \chi)$. Prove that $J(\chi, \chi) = a + b\omega$ where $\omega = \frac{-1+\sqrt{-3}}{2}$ and $a, b \in \mathbb{Z}$. Conclude that $A = 2a - b$ and $a^2 - ab + b^2 = p$. Furthermore, prove that if we let $B = b/3$ then

$$4p = A^2 + 27B^2.$$

(One can prove that B is an integer – see comments before Theorem 14.5.4.)

(67) For $p = 7, 13$, calculate by hand the points on $x^3 + y^3 = 1$ over $\mathbb{F}_p$, as well as the A appearing in Exercise 66, and verify Gauss's Theorem.

Using Gauss's theorem, find the number of solutions for the equation $x^3 + y^3 = 1$ for $p = 97$.

(68) Does it ever happen for $p \equiv 1 \pmod 3$ that $N(x^3 + y^3 = 1) = p$? What about $N(x^3 + y^3 = 1) = p - 1$? Suppose that p and $p - 2$ are primes, can it happen that the number of solutions to $x^3 + y^3 = 1 \bmod p$ is the same as the number of solutions mod $p - 2$? Explain how to find large p for which A, appearing in Exercise 66, is close to $2\sqrt{p}$ (and thus $N(x^3 + y^3 = 1)$ is very close to the maximum possible number of points[43] allowed by the Hasse bound $p + 2\sqrt{p}$).

[43] One can show that $x^3 + y^3 = z^3$ is an elliptic curve. This is easier if one allows a more general definition of elliptic curves that doesn't insist on describing them by equations of the form $y^2 = f(x)$.

(69) Let $d_i = \gcd(\ell_i, q-1)$, $a_i \in \mathbb{F}_q^\times$. Prove the equality for the number of solutions in $\mathbb{F}_q$:

$$N(a_1 x_1^{\ell_1} + \cdots + a_r x_r^{\ell_r} = b) = N(a_1 x_1^{d_1} + \cdots + a_r x_r^{d_r} = b).$$

(70) Prove that

$$\left| N^{\mathrm{proj}}(a_0 y_0^m + \cdots + a_n y_n^m = 0) - \sharp \mathbb{P}^{n-1}(\mathbb{F}_q) \right| \leq f(m) \cdot q^{\frac{n-1}{2}}.$$

where

$$f(m) = \frac{1}{m}((m-1)^{n+1} + (-1)^{n+1}(m-1)).$$

(71) ★ Prove that ζ_V provided in Theorem 15.0.1 satisfies the Weil conjectures.

(72) Give an explicit formula for $N^{\mathrm{proj}}(x_0^2 + \cdots + x_n^2 = 0)$ in $\mathbb{F}_q$.

(73) Give an explicit formula for $N^{\mathrm{proj}}(a_0 x_0^2 + \cdots + a_n x_n^2 = 0)$ in $\mathbb{F}_q$ ($a_i \in \mathbb{F}_q^\times, \forall i$).

(74) Let λ be the Legendre character. Let α be any non-trivial character of $\mathbb{F}_p^\times$. Prove that

$$J(\lambda, \alpha) = \sum_{t \in \mathbb{F}} \alpha(1 - t^2).$$

(Hint: use $N(x^2 = a) = 1 + \lambda(a)$.)

(75) Consider the equation $y^2 = x^3 + a$, where $a \in \mathbb{F}_p^\times$ is fixed and $p > 3$. Find an expression for $N(y^2 = x^3 + a)$. This expression will involve $J(\lambda, \alpha)$ where α is a cubic character. How does this compare with the expression for the zeta function of the projectivized curve $y^2 z = x^3 + az^3$?

(76) Let $p > 2$ be a prime and consider an equation of the form

$$C\colon \quad y^2 = f(x),$$

where f is a separable polynomial in $\mathbb{F}_p[x]$ of the degree $2g+1$.

- Prove that this is a non-singular curve in $\mathbb{A}^2$.
- Check that the corresponding projective curve in $\mathbb{P}^2$, obtained by homogenizing $y^2 - f(x)$ is singular if $g > 1$. However, one can show that there is a projective non-singular curve $\tilde{C}$ (living in some higher dimensional projective space) that contains C and such that $\tilde{C} \setminus C$ consists of a single point which is moreover defined over $\mathbb{F}_p$. The genus of $\tilde{C}$ is g and that implies that

  $$\zeta_{\tilde{C}}(T) = \frac{P_1(T)}{(1-T)(1-pT)},$$

  where $P_1 \in \mathbb{Z}[T]$ is a polynomial of degree $2g$ and constant coefficient 1.
  Assuming all that show that

  $$\sharp C(\mathbb{F}_p) = p + \sum_{t \in \mathbb{F}_p} \left(\frac{f(t)}{p} \right),$$

  and deduce the estimate due to Burgess

  $$\left| \sum_{t \in \mathbb{F}_p} \left(\frac{f(t)}{p} \right) \right| \leq 2g\sqrt{p}.$$

(77) Let $a, b, c \in \mathbb{F}_p$, $p > 2$,$a \neq 0, b^2 - 4ac \neq 0$. Determine the zeta function of the affine equation:

$$ax^2 + bxy + cy^2 = 1.$$

(78) ★ Let $\mathscr{L} \subset \mathbb{R}^n$ be a free abelian group of rank n. Then, $\mathscr{L}$ is a lattice if and only if $\mathscr{L}$ contains a basis of $\mathbb{R}^n$.

(79) Let $d > 0$ be an integer which is not a square. Consider the ring

$$\mathbb{Z}[\sqrt{d}] = \{a + b\sqrt{d} : a, b \in \mathbb{Z}\}.$$

Prove that the map

$$a + b\sqrt{d} \mapsto (a + b\sqrt{d}, a - b\sqrt{d}) \in \mathbb{R}^2,$$

realizes $\mathbb{Z}[\sqrt{d}]$ as a lattice in $\mathbb{R}^2$. What is the intersection of this lattice with the circle $x^2 + y^2 = 1$? the hyperbola $xy = 1$?

(80) Let $\mathscr{L}$ be a lattice with a generator matrix A. Show that $\mathscr{L}$ is integral if and only if ${}^tAA$ has integer entries.

(81) Let A be a generator matrix for a lattice $\mathscr{L}$. Prove that ${}^tA^{-1}$ is a generator matrix for the dual lattice $\mathscr{L}^\perp$. Conclude that $(\mathscr{L}^\perp)^\perp = \mathscr{L}$.

(82) Let $\mathscr{L}$ be an integral lattice and let $\mathscr{L}_1 \subseteq \mathscr{L}$ be a sub lattice. Prove that $\mathscr{L}_1^\perp \supseteq \mathscr{L}^\perp$ and $[\mathscr{L}_1^\perp : \mathscr{L}^\perp] = [\mathscr{L} : \mathscr{L}_1]$. (Hint: if $[\mathscr{L} : \mathscr{L}_1] = m$ then $\mathrm{covol}(\mathscr{L}_1) = m \cdot \mathrm{covol}(\mathscr{L})$.)

(83) Calculate the discriminant and the dual lattice of the following lattices.
- (a) $\mathscr{L} = \mathbb{Z}^n$.
- (b) Let m be a positive integer, $\mathscr{L} = \{(a_1, \ldots, a_n) \in \mathbb{Z}^n : \sum_{i=1}^n a_i \equiv 0 \pmod{m}\}$. (The lattices one gets for $m = 2$ are called the D_n lattices.)
- (c) $\mathscr{L} =$ the hexagonal lattice.
- (d) Let $d > 1$ be a square free integer. Consider the ring $\mathbb{Z}[\sqrt{-d}]$. Under the identification of $\mathbb{C}$ with $\mathbb{R}^2$ it becomes a lattice $\mathscr{L} \subset \mathbb{R}^2$. Write a generator matrix and a Gram matrix for $\mathscr{L}$; find the discriminant and the dual lattice. Is this an integral lattice?

(84) ★ A lattice is called **self-dual**, or **unimodular**, if $\mathscr{L} = \mathscr{L}^\perp$. Show that the only unimodular lattice in $\mathbb{R}^2$, up to isometry, is $\mathbb{Z}^2$.

(85) Consider the quadratic forms $q(x) = x^2 + y^2$ and $q(x) = x^2 - xy + y^2$. Find lattices in $\mathbb{R}^2$ with these quadratic forms (namely, that they have a Gram matrix with associated quadratic form given by q).

(86) Let $(x, y, z) \in \mathbb{R}^3$ and consider the abelian group generated by $(1, 0, 0), (0, 1, 0)$ and (x, y, z). Namely, $\mathbb{Z}(1, 0, 0) + \mathbb{Z}(0, 1, 0) + \mathbb{Z}(x, y, z)$. What are the conditions for it to be free of rank 3? What are the conditions for it to be a lattice? What are the conditions for it to be an integral lattice? a self-dual lattice?

(87) Let $\mathscr{L}$ be a lattice in $\mathbb{R}^n$. Prove that $\omega_n \geq \left(\frac{2}{\sqrt{n}}\right)^n$ and deduce that $\mathscr{L}$ contains a non-zero vector of length at most

$$\sqrt{n} \cdot (\mathrm{covol}(\mathscr{L}))^{1/n}.$$

(88) Prove that if $p > 2$ is a prime, $p \equiv 1 \pmod{3}$, then p is of the form $x^2 + 3y^2$.

(89) Prove that if $p > 2$ is a prime, $p \equiv 1 \pmod{8}$, then p is of the form $x^2 + 2y^2$.

(90) Let $\mathbb{H}$ denote the Hamilton quaternions (over $\mathbb{R}$). Prove that the map $\mathbb{H} \to \mathbb{H}, z \mapsto z^* := \mathrm{Tr}(z) - z$ is an anti-involution. Namely, it satisfies

$$(z_1 + z_2)^* = z_1^* + z_2^*, \quad (z_1 z_2)^* = z_2^* z_1^*.$$

Prove also that $\mathrm{Nm}(z) = zz^*$. (Suggestion: think in terms of matrices.)

(91) Prove that $\mathbb{H}$ is a non-commutative division ring (for any $x \neq 0$ there is a y such that $xy = yx = 1$). One reason this is interesting is that there is no commutative division ring of dimension 4 over $\mathbb{R}$, but here we see that there is a non-commutative division ring.

(92) The Hurwitz quaternions is the subset of Hamilton quaternions $\mathbb{H}$ given by

$$\mathbb{Z}[i,j,\frac{1+i+j+k}{2}] = \left\{a+bi+cj+d\ \frac{1+i+j+k}{2} : a,b,c,d \in \mathbb{Z}\right\}$$
$$= \left\{a+bi+cj+dk \in \mathbb{H} : a,b,c,d \in \mathbb{Z} \text{ or } a,b,c,d \in \mathbb{Z}+\tfrac{1}{2}\right\}.$$

Prove that the Hurwitz quaternions form a subring of $\mathbb{H}$. Prove that Nm is still integer valued on $\mathbb{Z}[i,j,\frac{1+i+j+k}{2}]$.

(93) Prove the following generalization of Dirichlet's theorem, by constructing a suitable convex symmetric set in $\mathbb{R}^{d+1}$. Let $\theta_1,\dots,\theta_d$ be real numbers and let $Q \in \mathbb{N}^+$. There there are integers $p_1,\dots,p_d,q$, not all zero, such that $0 \le q \le Q$ and

$$|q\theta_i - p_i| \le \frac{1}{Q^{1/d}},\ \ \forall i.$$

(94) Let $z = (z_1,\dots,z_n)$ be a primitive vector in $\mathbb{Z}^n$. Prove that there is a matrix $M \in \mathrm{GL}_n(\mathbb{Z})$ whose first column is ${}^t(z_1,\dots,z_n)$. (This is equivalent to showing that z can be completed to a basis of $\mathbb{Z}^n$. Consider $\mathbb{Z}^n/\mathbb{Z}z$ and prove first that it is a free abelian group of rank $n-1$.)

(95) ★ Derive a theorem similar to Theorem 20.5.2, but for the norm

$$\|(x_1,\dots,x_n)\|_1 = |x_1| + \cdots + |x_n|.$$

Namely, in this case we are trying to minimize the total amount of memory needed to store the solution in its entirety and not minimize every x_i separately.

(96) Find the successive minima and covol($\mathscr{L}$) for the following lattices. Write numerically the quantities in Minkowski's lattice point and successive minima theorems.
(a) $\mathscr{L} = \mathbb{Z} \oplus \mathbb{Z}i$, identified with $\mathbb{Z}^2$.
(b) $\mathscr{L} = \mathbb{Z} \oplus \mathbb{Z}\omega$, $\omega = \frac{-1+\sqrt{-3}}{2} \subset \mathbb{C} \cong \mathbb{R}^2$.
(c) $\mathscr{L} = \mathrm{Span}_{\mathbb{Z}}((1,0),(r_1,r_2))$, where r_1, r_2 are non-negative real numbers and $r_2 > 1$. (For μ_2 find only an approximation.)

(97) **The D_n lattices.** The D_n lattice in $\mathbb{R}^n$ is defined as

$$D_n = \{(x_1,\dots,x_n) \in \mathbb{Z}^n : \sum_{i=1}^n x_i \equiv 0 \pmod 2\}.$$

Compare Exercise 96. Find its successive minima. Find also $\sharp\{x \in D_n : \|x\| = \mu_1(D_n)\}$.

(98) Consider the lattice $A_n = \{(z_0,\dots,z_n) \in \mathbb{Z}^{n+1} : \sum_{i=0}^n z_i = 0\}$. Prove that covol$(A_n) = n+1$. Prove that $\mu_i(A_n) = \sqrt{2}$ for all i. Find

$$\sharp\{x \in A_n : \|x\| = \mu_1\}.$$

(Compare this with the lattice $\mathbb{Z}^n$ that also has all its successive minima equal but for which $\sharp\{x \in \mathbb{Z}^n : \|x\| = \mu_1\} = 2n$.)

(99) Is it true or not that A_3, properly rescaled, is isometric to D_3? What about A_4 and D_4?

(100) Show that the lattice A_2 can be identified with a lattice in $\mathbb{R}^2$ that is, up to scaling and perhaps rotation, the hexagonal lattice.

(101) Find a generator matrix for A_n^*, the dual lattice to the lattice A_n. Determine covol(A_n^*) and prove that $\mu_1(A_n^*) = \sqrt{n/(n+1)}$ and that it is achieved $2n+2$ times if $n \ge 2$ and 2 times if $n = 1$.

(102) ★ Write down the Gram matrix and calculate covol(E_6). Show that $\mu_1(E_6) = \sqrt{2}$ and it is achieved by 72 vectors – this is called the **kissing number** of the lattice. (Note that for $\mathbb{Z}^6$ this number is 12, for A_6 it is 42 and for D_6 it is 60.) The lattice E_6 is known to achieve the highest kissing number among all lattices in $\mathbb{R}^6$, D_4 and D_5 hold the record for their dimension, and A_2 and A_3 for theirs.

(103) For which n, $\Delta(\mathbb{Z}^n) < 2^{-n}$? What explanation is offered by the proof of Proposition 23.1.3?

(104) ★Prove that for every lattice $\mathscr{L}$,

$$R(\mathscr{L}) \leq \sqrt{n} \times \frac{1}{2}\mu_n(\mathscr{L}).$$

(105) In light of Exercise 104, $\mathbb{Z}^n$ has the worst covering radius, in the sense that $\frac{R(\mathbb{Z}^n)}{\frac{1}{2}\mu_n(\mathbb{Z}^n)}$ attains the maximum possible. Find this ratio for the hexagonal lattice and the plane lattices $\mathbb{Z}[\sqrt{-d}]$, where $d > 0$ is an integer and we identify $\mathbb{C}$ with $\mathbb{R}^2$.

(106) Prove that if C is a (n,k,d)-code then

$$d \leq n - k + 1.$$

(107) Determine $(n(C^e), k(C^e), d(C^e))$ in terms of $(n(C), k(C), d(C))$. Determine W_{C^e} in terms of W_C. Determine $(C^e)^\perp$ in terms of $C^\perp$.

(108) A code is called **self-dual** if $C = C^\perp$. Prove that in this case n is even and $k(C) = n/2$. Prove also that every code word has even weight. Prove that

$$W_C(x,y) = W_C(y,x).$$

(This can be proven using the MacWilliams identity.)

(109) Prove that for a self-dual code C,

$$W_C(x,y) = W_C(\frac{x+y}{\sqrt{2}}, \frac{x-y}{\sqrt{2}}).$$

Let D be the group of matrices generated by

$$\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Prove that D is the dihedral group of 16 elements. Prove that if C is a self-dual code then $W_C(x,y)$ is invariant under the group D that acts on polynomials by $f(x,y) \mapsto f((x,y)A), A \in D$.

Prove that the polynomials

$$\phi_2 = x^2 + y^2, \quad \phi_8 = x^8 + 14x^4y^4 + y^8,$$

are invariant under D. It is a theorem of A. M. Gleason that for a self-dual code C, W_C is always a polynomial expression in ϕ_2 and ϕ_8.

(110) Let C_1, C_2 be codes. The code $C_1 \oplus C_2$ is defined as

$$\{(x,y) : x \in C_1, y \in C_2\}.$$

If C_i is an (n_i, k_i, d_i) code, what is the type of $C_1 \oplus C_2$? Prove that if C_i are both self-dual so is $C_1 \oplus C_2$. Prove that

$$W_{C_1 \oplus C_2}(x,y) = W_{C_1}(x,y)W_{C_2}(x,y).$$

Find all self-dual codes of dimension $2, 4, 6$ and their weight enumerator polynomials. How do your examples compare with Gleason's theorem?

(111) By considering the cyclic shifts of v_1 conclude that $\mathscr{H}_7$ has at least 7 code words of weight 3. Find a vector of weight 4 and use it to show that $\mathscr{H}_7$ has at least 7 code words of length 4. Show also that there is a code word of weight 7 (Hint: what polynomial will it correspond to?). Explain that this is enough to conclude that

$$W_{\mathscr{H}_7}(x,y) = x^7 + 7x^4y^3 + 7x^3y^4 + y^7.$$

In particular, deduce this way that the distance of $\mathscr{H}_7$ is 3.

(112) Prove that $\mathscr{H}_8 = \mathscr{H}_7^e$ is an $(8,4,4)$ self-dual code with

$$W_{\mathscr{H}_8}(x,y) = x^8 + 14x^4y^4 + y^8.$$

(113) Prove that a cyclic code C associated to $g(t)$ is self-dual, if and only if (in the notation of Theorem 24.4.7) $g(t) = f(t)$, and necessarily n is even. Prove that if $n = 2r$ then $1 + t^r$ defines a cyclic self-dual code.

(114) Find all self-dual cyclic codes of length $2, 4, 6, 8, 10$.

(115) ★ Find all self-dual cyclic codes of length 14.

(116) One thought regarding error-correcting is that we may just send every block of size k twice. Consider this for the Golay code. This idea suggests that instead of using the Golay code which is of length 23, we can use the code C of dimension 24 which is a variant on a repetition code.

$$C = \{(x,x) : x \in \mathbb{F}_2^{12}\} \subset \mathbb{F}_2^{24}.$$

Discuss the advantages and disadvantages of this idea.

(117) The Golay code $\mathscr{G}_{23}$ turns out to be also a special case of a quadratic residue code (as is the Hamming code $\mathscr{H}_7$). We don't enter into the general theory of such codes here, but it implies that the Golay code is also the cyclic code generated by

$$f(t) = t + t^2 + t^3 + t^4 + t^6 + t^8 + t^9 + t^{12} + t^{13} + t^{16} + t^{18}.$$

(The meaning of that is that the ideal generated by $f(t)$ in $\mathbb{F}_2[t]/(t^n - 1)$ is the same as the one used to define the Golay code.) It also implies that the Hamming code is also generated by

$$t + t^2 + t^4.$$

(118) Verify Theorem 25.1.1 for the codes Z, U, R, P and E_8.

(119) Prove the identity $\theta_{\mathscr{L}_1 \oplus \mathscr{L}_2} = \theta_{\mathscr{L}_1}\theta_{\mathscr{L}_2}$. Prove that the coefficient of q^m in $(\theta_{\mathbb{Z}})^4$ is positive for every $m \geq 0$.

(120) Write an expression for Θ_{E_8} in terms of θ_2 and θ_3. Use it to find the first minimum of E_8 and verify that its kissing number is 240. Using the generator matrix for E_8 now determine all the successive minima of E_8.

(121) Prove that the lattice E_8 is a unimodular lattice. Namely, E_8 is self-dual and $\mathrm{covol}(E_8) = 1$. Prove that the same is true for $\mathbb{Z}^8$. Prove, using Theorem 25.1.1, that the kissing number of E_8 is 240, while for $\mathbb{Z}^8$ it is 16. This again illustrate how dramatically better the E_8 -packing is in comparison to the square packing provided by $\mathbb{Z}^8$.

(122) Recall that $D_n = \{(x_1, \ldots, x_n) \in \mathbb{Z}^n : \sum x_i \equiv 0 \pmod 2\}$. Let $[\tfrac{1}{2}] = (\tfrac{1}{2}, \tfrac{1}{2}, \ldots, \tfrac{1}{2}) \in \mathbb{R}^n$. Let

$$D_n^+ = D_n \coprod ([\tfrac{1}{2}] + D_n).$$

(a) Prove that D_n^+ is a lattice if and only if n is even.
(b) Prove that D_n^+ is an integral lattice if and only if $4|n$.
(c) Prove that D_n^+ is even if and only if $8|n$.
(d) Prove that $\mathrm{covol}(D_n^+) = 1$.
(e) For n even, prove that $\mu_1(D_n) = \mu_1(D_n^+)$ and calculate $\delta(D_n^+)$.

(123) ☺ **The problem of the square pyramid.** It was conjectured by E. Lucas in 1875 and finally proven by G. N. Watson in 1918,[44] that the only cases where a sum of squares $1^2+2^2+\cdots+n^2$ is a square of an integer z are the cases where $(n,z)\in\{(1,1),(24,70)\}$. The origin of the problem is that when cannon balls are laid in the form of a square pyramid, with 1 ball at the top, 4 balls in the second layer, 9 points in the third layer and so on. The number of balls is $1^2+2^2+\cdots+n^2$ and so one asks if the pyramid is a square.

This is a remarkably deep problem. The question is what are the positive integer solutions to the equation

$$z^2=\frac{n(n+1)(2n+1)}{6}.$$

Multiply the equation by 24 and put $x=2n, y=2z$ to reduce the problem to finding integral solutions to the equation $6y^2=x(x+1)(x+2)$, and putting $u=x+1$ to finding integral solutions the equation

$$6y^2=u^3-u.$$

The original solutions $[n,z]$ correspond to the solutions $[u,y]=[3,2],[49,140]$, to the last equation. Clearly there are other integral solutions. For example, $[0,0],[1,0],[2,\pm1]$.

The equation $6y^2=u^3-u$ defines an elliptic curve E over $\mathbb{Q}$. It turns out that there is an abelian group law on the set of rational points of E; in fact, this is true for any elliptic curve $dy^2=u^3+au+b$ over $\mathbb{Q}$ (the zero point 0_E is an ideal point "at infinity", visible as the point $[0:1:0]$ when completing the elliptic curve to a projective curve $dy^2v=u^3+auv^2+bv^3$ in the projective plane with coordinates u,y,v). The celebrated Mordell-Weil Theorem says that the group $E(\mathbb{Q})$, is finitely generated abelian group, hence isomorphic to $\mathbb{Z}^r\oplus T$ where r is called the rank of E over $\mathbb{Q}$, and T is a finite abelian group, the torsion group of $E(\mathbb{Q})$.

For the elliptic curve $6y^2=u^3-u$, we change coordinates once more by putting $y=Y/6^2, u=U/6$ to find the equation $Y^2=U^3-36*U$, which is in a form suitable to be tested by PARI. The points $[u,y]=[3,2],[49,140]$, correspond to the points $[18,72],[294,5040]$ in the coordinates $[U,Y]$.

The command `E = ellinit([0, 0, 0, -36])` creates the elliptic curve $Y^2=U^3-36*U$ in PARI. Then `ellanalyticrank(E)` calculates the rank of $E(\mathbb{Q})$ and returns in our case the value 1. The command `elltors(E)` returns the values

```
[4, [2, 2], [[-6, 0], [0, 0]]],
```

which means the torsion group has 4 elements, is isomorphic to $\mathbb{Z}/2\mathbb{Z}\times\mathbb{Z}/2\mathbb{Z}$ and is generated by the two points $[-6,0],[0,0]$. To find all torsion points we can add these two points using `elladd(E, [-6, 0], [0, 0])` and find the additional point $[6,0]$. Thus,

$$E^{\text{tors}}(\mathbb{Q})=\{0_E,[-6,0],[0,0],[6,0]\}.$$

Another powerful theorem due to Gross and Zagier tells us that when the rank of E is 1, as it is in our case, then a generator can be found by the method of Heegner points. The command

[44] G. N. Watson, "The problem of the square pyramid", *Messenger of Math.* XLYIII (1918), 1 - 22.

`ellheegner(E)` calculates this point, which is $[12, 36]$, and so

$$E(\mathbb{Q}) = \mathbb{Z} \cdot [12, 36] + E^{\mathrm{tors}}(\mathbb{Q}).$$

One must remember that, for example, $3 \cdot [12, 36]$ means adding the point $[12, 36]$ to itself 3 times using the elliptic curve group law. You can do that using the command `ellmul(E, [12, 36], 3)` and find the point $[16428/529, -2065932/12167]$.

From this perspective, the difficulty of Lucas's problem is that there are infinitely many rational solutions on the curve $Y^2 = U^3 - 36 * U$. In fact, if (n, z) is an integral solution to the original problem, the point on $Y^2 = U^3 - 36 * U$ is given as $[6(2n+1), 72z]$ and, so, in effect we are looking for positive solutions (U, Y) such that $72|Y$ and $U/6$ is an odd integer. This is not an easy problem. For a modern treatment of a class of similar problems, and literature review, see M. Bennet, "Lucas' square pyramid problem revisited". Acta Arith. 105 (2002), no. 4, 341–347. In particular the solution by I. Cucurezeanu in "An Elementary Solution of Lucas' Problem", Journal of Number Theory 44 (1993), 9-12, is attractive in its simple methods; as the abstract states "By the method of infinite descent, Pell's equation, and the quadratic reciprocity law, it is proved that the equation $x(x+1)(2x+1) = 6y^2$ has the only nontrivial integer solution $x = 24, y = 70$."

(124) ☺ Let $n \geq 2$. Let A, B be positive integers. Find a sufficient condition that a positive solution (x, y) for the equation

$$x^n - Ay^n = B,$$

arises as a convergent x/y for $\alpha := \sqrt[n]{A}$. Use the factorization

$$x^n - Ay^n = (x - \alpha y)(x^{n-1} + \alpha x^{n-2} y + \alpha^2 x^{n-2} y^2 + \cdots + \alpha^{n-1} y^{n-1}).$$

This is a special kind of a so-called Thue equation. It is know that any Thue equation has only finitely many integer solutions.

(125) ☺ Let $\chi \in \mathbb{X}_q$. Calculate $\sum_{a \in \mathbb{F}_q} \mathfrak{g}_a(\chi)$.