Diophantine Analysis

Jörn Steuding

Winter 2002/03
Johann Wolfgang Goethe-Universität Frankfurt

This course gives an introduction to the theory of *Diophantine Approximations* with a special emphasis on its impact to the field of *Diophantine Equations*. It relies mainly on the nicely written books [9], [29] as well as the more general introduction [8] and the classics [21], [35]; historical details can be found in [11], [48] and on $\mathsf{www.history.mcs.st-andrews.ac.uk}$. Finally, we refer to [4] for modern perspectives in this amazing and growing field.

We will use only standard notation. Special knowledge which is beyond the scope of first courses in mathematics is (with some small exceptions) not necessary. The reader can find some easy and some difficult exercises in the text which may help to get familar with the topic.

The author is very grateful to Iqbal Lahseb, Ernesto Girondo, Thomas Müller and Matthias Völz for several remarks and corrections.

1 Introduction: three basic principles

Diophantus of Alexandria was a greek mathematician who lived around 250 A.D., but virtually nothing more about his life is known than that what was written down on his tombstone:

> *God granted him to be a boy for the sixth part of his life,*
> *and adding a twelfth part to this, He clothed his cheeks with down.*
> *He lit him the light of wedlock after a seventh part,*
> *and five years after his marriage He granted him a son.*
> *Alas! late-born wretched child; after attaining*
> *the measure of half his father's life, chill Fate took him.*
> *After consoling his grief by this science of numbers*
> *for four years he ended his life.* (cf. [52], p.55)

How old was Diophantus when he died? This riddle is an example for the kind of problems in which Diophantus was interested in. He wrote the influential monography *Arithmetica* which inspired Fermat (1607/08?-1665) to write down

It is impossible for a cube to be written as a sum of two cubes
or a fourth power to be written as the sum of two fourth powers or,
in general, for any number which is a power greater than the second
to be written as a sum of two like powers.
I have a truly marvellous demonstration of this proposition
which this margin is too narrow to contain. (cf. [52], p.66)

in his copy of Diophantus' book, exactly there where the corresponding question for squares is considered. In the modern language of algebra Fermat claimed to have a proof that *all solutions of the equation*

$$X^n + Y^n = Z^n \tag{1}$$

in integers x, y, z *are trivial, i.e.* $xyz = 0$*, whenever* $n \geq 3$. Fermat never published a proof and, by the succcessless quest for a solution of **Fermat's last theorem** over centuries, mathematicians started to believe that Fermat actually had no proof. However, no counterexample was found. Only recently Wiles (supported by Taylor and the earlier works of many others) found a proof for Fermat's last theorem (see [52] for the amazing story of this problem and its final solution). With our modern background in mathematics one may ask why the ancient Greek did not consider the Fermat equation in its generality but only the quadratic case. Their point of view was inspired by at most three-dimensional geometry and only in the late works of Greek mathematicians higher powers occur. They also had an advanced knowledge on divisibility and prime numbers but no idea about the unique prime factorization of the integers!

It is interesting that the exponent in the Fermat equation is crucial. As Diophantus explained in his book the equation (1) with exponent $n = 2$ has infinitely many (non-trivial) solutions, for example

$$3^2 + 4^2 = 5^2, \quad 5^2 + 12^2 = 13^2, \quad 8^2 + 15^2 = 17^2, \quad \dots .$$

These triples of integral solutions are called **Pythagorean triples** with regard to Pythagoras' theorem in geometry. However, the history of such triples dates at least back to the ancient Babylonians four millenia ago. It is conjectured that they used Pythagorean triples for constructing right angles. Pythagoras (572-492 B.C.) found not only a proof that such triples yield right angles but gave also an infinitude of primitive (coprime) Pythagorean triples by the identity

$$(2n+1)^2 + (2n^2+2n)^2 = (2n^2+2n+1)^2 \qquad \text{for} \quad n \in \mathbb{N}.$$

In the third century B.C. Euclid solved the problem of finding all solutions.

Exercise 1 *Show that all positive integral solutions* x, y, z *of*

$$X^2 + Y^2 = Z^2,$$

satisfying $(x,y)=1$ *and* $2|x$, *are given by*

$$x=2ab, \quad y=a^2-b^2, \quad z=a^2+b^2,$$

where a *and* b *are coprime integers of opposite parity and* $a>b$. *[Hint: each solution satisfies* $x^2=z^2-y^2=(z+y)(z-y)$*.]*

Note that this classification of the Pythagorean triples can be used to prove that the biquadratic case of (1) has only trivial solutions. Actually, Fermat combined this result with his *method of infinite descent* to prove that the slightly more general equation

$$X^4+Y^4=Z^2$$

has no positive integral solutions (see [21], §13.3).

We are living in a finite universe (there are only about 10^{80} atoms in our universe) which means that our world can be described using only rational numbers. However, for an *understanding* of the physical world as well as the abstract world of mathematics this is not enough! The same equation which Euclid studied so successfully led some hundred years before to one of the great breakthroughs in ancient Greek mathematics. Hippasus, a pupil of Pythagoras, discovered that the set of rational numbers is too small for the simple geometry of triangles and squares. In fact, he proved that the length of the diagonal of a unit square is irrational:

$$\sqrt{2}=\sqrt{1^2+1^2}\not\in\mathbb{Q}.$$

Nowadays this is taught at school, but for the Pythagoras school it was the death of their philosophy that all natural phenomenon could be explained in integers. It is said that Pythagoras sentenced Hippasus to death by drowning. This unkind act could not stop the mathematical progress. In order to solve polynomial equations and for the merits in analysis the mathematicians invented various *new* numbers:

$$\mathbb{N}\subset\mathbb{Z}\subset\mathbb{Q}\subset\mathbb{R}\subset\mathbb{C},$$

and even more as we will see in section 16; we refer the interested reader to the collection [12] of excellently written surveys on numbers of all kinds. However, for the first we shall only consider real numbers.

By Cantor's diagonal argument we know that *the set of real numbers* $\mathbb{R}$ *is uncountable whereas* $\mathbb{Q}$ *is countable.* Therefore, almost all real numbers are irrational. We shall give a more advanced example than the square root of two. One of the fundamental constants in analysis (and natural sciences) is the value of the exponential function at one:

$$e=\sum_{n=0}^{\infty}\frac{1}{n!}.$$

Theorem 1 *e is irrational.*

This result dates back to Euler in 1737, resp. Lambert in 1760; however their approach via *continued fractions* is rather difficult (we will return to this later). However, the following simple proof would have been possible in Euler's times.

Proof. Suppose the contrary, then there exist positive integers a, b such that $e = \frac{a}{b}$. If m is an integer $\geq b$ then b divides $m!$ and

$$\alpha = m!\left(e - \sum_{n=0}^{m} \frac{1}{n!}\right) = a\frac{m!}{b} - \sum_{n=0}^{m} \frac{m!}{n!}$$

is an integer. On the other side,

$$0 < \alpha = \sum_{n=m+1}^{\infty} \frac{m!}{n!} < \frac{1}{m+1}\sum_{k=0}^{\infty}\left(\frac{1}{m+1}\right)^k = \frac{1}{m},$$

which gives the contradiction. The theorem is proved. •

This proof reveals an important principle in the diophantine toolbox: *the series converges so fast that the limit cannot be of a restricted arithmetical nature!*

The question whether a given real number is irrational seems to be simple on the first view. Actually, this is a rather difficult problem. For instance, it is unknown whether the **Euler-Mascheroni constant**

$$\gamma = \lim_{N\to\infty}\left(\sum_{n=1}^{N} \frac{1}{n} - \log N\right)$$

is irrational. Another important constant is π, which we define as the smallest positive root of the sine function.

Theorem 2 *π is irrational.*

The first proof of the irrationality of π was given by Lambert in 1761, also by using continued fractions. The proof which we shall give now, as all other known proofs, is slightly more difficult than the given one for e. This holds also in other questions around these two fundamental numbers. It seems that e has somehow more *structure* than π. Our short but tricky proof is due to Niven [37].

Proof. We start with some preliminaries. For $n \in \mathbb{N}$ define

$$f_n(x) = \frac{1}{n!}x^n(1-x)^n. \tag{2}$$

Hence,

$$f_n(x) = \frac{1}{n!}\sum_{j=n}^{2n} c_j x^j \qquad \text{with} \quad c_j \in \mathbb{Z}$$

and

$$0 < f_n(x) < \frac{1}{n!} \qquad \text{for} \quad 0 < x < 1. \tag{3}$$

It is easy to find for the k-th derivative

$$(-1)^k f_n^{(k)}(1) = f_n^{(k)}(0) = \begin{cases} 0 & \text{if } \ 0 \leq k < n, \\ \frac{k!}{n!} c_k & \text{if } \ n \leq k \leq 2n; \end{cases} \tag{4}$$

here we used $f_n^{(k)}(x) = (-1)^k f_n^{(k)}(1-x)$ and the Taylor expansion (or dumb computation). Note that all these values are integers.

We shall even show that π^2 *is irrational.* Assume that $\pi^2 = \frac{a}{b}$ with positive integers a and b. We consider the polynomial

$$F_n(x) := b^n(\pi^{2n} f_n(x) - \pi^{2n-2} f_n^{(2)}(x) \pm \ldots + (-1)^n f_n^{(2n)}(x)).$$

Since $b^n \pi^{2k} = b^{n-k} a^k \in \mathbb{Z}$ for $0 \leq k \leq n$, it follows in view of (4) that $F_n(0), F_n(1) \in \mathbb{Z}$. A short calculation shows

$$(F_n'(x)\sin(\pi x) - \pi F_n(x)\cos(\pi x))' = \pi^2 a^n f_n(x) \sin(\pi x).$$

With regard to $\sin \pi = \sin 0 = 0$ this yields

$$\mathcal{I}_n := \pi a^n \int_0^1 f_n(x) \sin(\pi x) \, \mathrm{d}x = F_n(0) + F_n(1),$$

which is an integer. On the other side we get with regard to (3)

$$0 < \mathcal{I}_n < \pi \frac{a^n}{n!}.$$

Since the factorial $n!$ grows faster than a^n, the right hand side above can be made < 1 by choosing n sufficiently large. This gives the contradiction and the theorem is proved. •

Also this proof is very interesting: *a problem concerning the arithmetic nature of a given real number is solved by the construction of an appropriate sequence of polynomials with respect to its analytical behaviour!*

As we have seen above almost all real numbers are irrational, but how to deal with them in our finite world? Since $\mathbb{Q}$ *is dense in* $\mathbb{R}$ it is natural to search for rational approximations (but $\mathbb{Q}$ has also the disadvantage of beeing very *thin* in $\mathbb{R}$). For example, we can think about some machine which has to compute an approximation to the circumference of a circle with given radius. For that we have to construct some mechanism with gears which approximates the ratio $\pi : 1$, but this needs a *good* approximation of the irrational

$$\pi = 3.14159\,26535\ldots$$

by rational numbers. It is interesting to see how this problem was solved in former times:

- The Rhind Papyrus (≈ 1650 B.C.): $\pi \approx 4\left(\frac{8}{9}\right)^2 = 3.16045\ldots$;

- Old testament (≈ 1000 B.C.): $\pi \approx 3$;
- Archimedes (287-212 B.C.): $\pi \approx \frac{22}{7} = 3.14285\ldots$;
- Tsu Chung Chi (≈ 500 A.D.): $\pi \approx \frac{355}{113} = 3.14159\,29920\ldots$.

We shall not speak about the attempt of an american politician to attach π by law the value 3.2 in 1897; for this and more on π see [5]. More useful is the rhyme

Now I want a drink, alcoholic of course,
after the heavy lectures involving quantum mechanics!

Recently, Kanada and Takahashi computed π up to more than 206 billion digits. This precision is beyond any use in applications (the Planck constant 10^{-33} is the smallest unit in quantum mechanics) but interesting from a mathematical point of view. Since $\mathbb{Q}$ is dense in $\mathbb{R}$, for any real number there exist infintely many rational approximations. Moreover, we can approximate with any assigned degree of accuracy. But what are *good* and what are *bad* approximation among them or, what is the same, how *rapidly* can we approximate? A natural measure is the denominator. Let α be any real number, then we say that $\frac{p}{q} \in \mathbb{Q}$ with $q \geq 1$ is a **best approximation** to α if

$$|q\alpha - p| < |Q\alpha - P| \qquad \text{for all} \quad Q < q,$$

where $P, Q \in \mathbb{Z}$; necessarily a best approximation is a reduced fraction. In particular, a best approximation $\frac{p}{q}$ to α is the nearest rational number with denominator $\leq q$, but not vice versa! For example, the first best approximations to π are

$$\frac{3}{1}, \frac{22}{7}, \frac{333}{106}, \frac{355}{113}, \frac{1\,03993}{33102}, \ldots \to \pi.$$

It is remarkable that the fractions given by Archimedes and of Tsu Chung Chi are best approximations. How could they do that in these dark ages without computers? In general, finding good rational approximations to a given irrational number is a difficult task.

In honour of Diophantus we speak about

- **Diophantine Approximations** when we search for rational approximations to rational or irrational numbers;
- **Diophantine Equations** when we investigate polynomial equations on solutions in integers (or rationals).

As we shall see both areas are linked in various directions. For instance, we ask for solutions of the equation

$$38X - 15Y = 1 \tag{5}$$

in integers. One approach to answer this question offers Euclid's algorithm. By **division with remainder** *for any positive integers a, b with $b > a$ there exist integer q, r such that*

$$b = aq + r \qquad \textit{with} \quad 0 \leq r < a.$$

Now define $r_{-1} := b, r_0 := a$. Then, successive application of division with remainder yields the **Euclidean algorithm**:

$$\textsf{For} \quad j = 0, 1, \ldots \quad \textsf{do} \quad r_{j-1} = q_{j+1} r_j + r_{j+1} \qquad \textsf{with} \quad 0 \leq r_{j+1} < r_j. \tag{6}$$

Since the sequence of remainders is strictly decreasing, the algorithm finishes after finitely many, say n steps and, by simplest divisibility properties, it turns out that the last non-vanishing remainder r_n is equal to the greatest common divisor of a and b.

Exercise 2 *Show that the running time (i.e. the number of steps) in the Euclidean algorithm is*

$$n \leq \frac{2}{\log 2} \log(b + 1).$$

Can you describe the smallest numbers a, b for which the algorithm needs exactly n steps? Can you improve the previous estimate?

From an algorithmical point of view it is of advantage to use a modified Euclidean algorithm which works with the nearest integer and not with the next larger integer. However, for our purpose the above stated version is sufficient. Reading the Euclidean algorithm backwards, we get a concrete integral solution of the linear equation

$$bX - aY = (a, b), \tag{7}$$

say x_0, y_0. It is easily seen that then all integral solutions to the latter diophantine equation are given by

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} + \frac{1}{(a, b)} \begin{pmatrix} a \\ b \end{pmatrix} \cdot \mathbb{Z}. \tag{8}$$

From here it is only a small step to

Theorem 3 *The linear diophantine equation $bX - aY = c$ with integers a, b, c is solvable if and only if $(a, b)|c$.*

We return to our example. It may be a little bit surprising to see that the solutions of (5) yield *good* approximations to $\frac{15}{38}$ and vive versa:

$$\ldots, \quad \mathbf{38} \cdot (-13) - \mathbf{15} \cdot (-33) = 1, \quad \mathbf{38} \cdot 2 - \mathbf{15} \cdot 5 = 1, \quad \mathbf{38} \cdot 17 - \mathbf{15} \cdot 43 = 1, \ \ldots$$

vs.

$$\frac{x}{y} = \frac{2}{5}, \frac{13}{33}, \frac{17}{43} \ldots \to \frac{\mathbf{15}}{\mathbf{38}};$$

note that the solutions x, y with $|y| < 38$ give even the best approximations $\frac{x}{y}$ to $\frac{15}{38}$. However, this is not too surprising. We may rewrite each solution of our modified equation as

$$\frac{x}{y} = \frac{15}{38} + \frac{1}{38y},$$

and since the second term is rather small, we see a certain relation between approximations to $\frac{15}{38}$ and our integral solutions.

Exercise 3 *Prove that the best approximations $\frac{x}{y}$ to any given $\frac{a}{b} \in \mathbb{Q}$ exactly correspond to those solutions $x, y \in \mathbb{Z}$ of (7) with $|y| < b$.*

With regard to our previous observations we can solve the linear diophantine equation (7) by searching for the best approximations to $\frac{a}{b}$. This was first discovered by Aryabhata around 550 A.D. However, this example is only the top of an iceberg: *certain diophantine equations can be investigated by studying a related problem in the theory of diophantine approximations!*

2 Dirichlet's approximation theorem

It is much more interesting to approximate irrational numbers than rationals. In 1842 Dirichlet proved the first but fundamental approximation theorem.

Theorem 4 *If α is irrational, then there exist infinitely many rational numbers $\frac{p}{q}$ such that*

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{q^2}; \tag{9}$$

this property characterizes irrational numbers, i.e., a rational α allows only finitely many approximations $\frac{p}{q}$ with (9).

In particular, there exist infinitely many best approximations to any given irrational number (in contrast to rationals; see Exercise 3).

The following original proof relies on the **pigeonhole principle** which states that *if $n+1$ objects are distributed to n boxes, then at least one box contains at least two objects* (which is easily seen by a contradiction argument or induction).

Proof. We define the **integral part** and the **fractional part** of a real number x by

$$[x] = \max\{z \in \mathbb{Z} : z \leq x\} \qquad \text{and} \qquad \{x\} = x - [x],$$

respectively. Let Q be a positive integer. The numbers

$$0, \{\alpha\}, \{2\alpha\}, \ldots, \{Q\alpha\}$$

define $Q+1$ points distributed among the Q disjoint intervals

$$\frac{j-1}{Q} \leq x < \frac{j}{Q} \qquad \text{for} \quad j = 1, \ldots Q.$$

By the pigeonhole principle there has to be at least one interval which contains at least two numbers $\{k\alpha\} > \{\ell\alpha\}$, say, with $0 \leq k, \ell \leq Q$. It follows that

$$\begin{aligned} 0 \leq \{k\alpha\} - \{\ell\alpha\} &= k\alpha - [k\alpha] - \ell\alpha + [\ell\alpha] \\ &= \{(k-\ell)\alpha\} + [(k-\ell)\alpha] + \underbrace{[\ell\alpha] - [k\alpha]}_{\in \mathbb{Z}} < \frac{1}{Q}. \end{aligned}$$

Obviously, the integral parts add up to zero. Hence, setting $q = k - \ell$ we obtain

$$\{q\alpha\} = \{k\alpha\} - \{\ell\alpha\} < \frac{1}{Q}.$$

With $p := [q\alpha]$ it follows that

$$\left|\alpha - \frac{p}{q}\right| = \frac{|q\alpha - p|}{q} < \frac{\{\alpha\}}{qQ} < \frac{1}{qQ}, \tag{10}$$

which implies the estimate (9) (since $q < Q$).

Now assume that α is rational, say $\alpha = \frac{a}{b}$. If $\alpha \neq \frac{p}{q}$, then

$$\left|\alpha - \frac{p}{q}\right| = \frac{|aq - bp|}{bq} \geq \frac{1}{bq} \tag{11}$$

and (9) involves $q < b$, which proves that there are only finitely many $\frac{p}{q}$ with (9).

Now suppose that α is irrational and that there exist only finitely many solutions $\frac{p_1}{q_1}, \ldots, \frac{p_n}{q_n}$ to (9). Since $\alpha \notin \mathbb{Q}$, we can find a Q such that

$$\left|\alpha - \frac{p_j}{q_j}\right| > \frac{1}{Q} \qquad \text{for} \quad j = 1, \ldots, n,$$

contradicting (10). The theorem is proved. •

Notice that by this proof it is impossible to compute approximations without big computational effort.

Dirichlet's approximation theorem can be extended to a first irrationality criterium:

Exercise 4 *Prove that if there are infinitely many coprime solutions p, q of*

$$|q\alpha - p| < q^{-\delta}$$

with a fixed $\delta > 0$, then α is irrational. [Hint: using (11) show that for rational α the sets of possible denominators q and of possible numerators p are finite.]
Deduce the irrationality of

$$\sum_{n=1}^{\infty} 2^{-n!}.$$

[The idea behind should be compared with our proof of the irrationality of e.]

More irrationality criteria of this type can be found in [8], §5.

Dirichlet's proof yields also an extension to simultaneous approximation problems.

Exercise 5 *Prove that if* $\alpha_1, \ldots, \alpha_n$ *are arbitray real numbers, then the system of inequalities*

$$\left|\alpha_j - \frac{p_j}{q}\right| < q^{-1-\frac{1}{n}} \qquad \textit{for} \quad j = 1, \ldots, n$$

has at least one solution $p_1, \ldots p_n, q \in \mathbb{Z}$*; if at least one* α_j *is irrational, then it has an infinitude of solutions.*

We say that α is **approximable by rationals of order** κ if there exists a positive constant $c(\alpha)$, depending only on α, such that

$$\left|\alpha - \frac{p}{q}\right| < \frac{c(\alpha)}{q^{\kappa}}$$

has an infinity of solutions $\frac{p}{q}$. In a sense, κ indicates the speed of convergence of the sequence $\frac{p}{q}$ to α. With view to our previous observations we see that

- any rational α is approximable of order 1, and to no higher order (by Exercises 3 and 5);
- any irrational α is at least approximable of order 2 (by Theorem 4).

It is a natural question to ask for improvements of Dirichlet's theorem. Khintchine proved in the 1930s a remarkable result which states that the set of numbers which may allow a stronger approximation has **measure null**, i.e given any positive ε, the set can be covered by a countable number of intervals of total length $< \varepsilon$.

Theorem 5 *Suppose that* ψ *is a positive function such that*

$$\sum_{q=1}^{\infty} \psi(q)$$

converges. Then for almost all α *(i.e. with exceptions which have measure* 0*), there is only a finite number of solutions* $p, q \in \mathbb{Z}$ *to the inequality*

$$|q\alpha - p| < \psi(q). \tag{12}$$

The proof relies on Dirichlet's idea of locating the real number α in question in small intervals.

Proof. Given $\varepsilon > 0$, we can find an integer Q such that

$$\sum_{q \geq Q} \psi(q) < \frac{\varepsilon}{2}.$$

Now consider those α for which the inequality (12) has infinitely many solutions. Now for each $q \geq Q$ consider the intervals of radius $\frac{\psi(q)}{q}$ surrounding the rational numbers $\frac{0}{q}, \frac{1}{q}, \ldots, \frac{q-1}{q}$. Consequently, each α will lie in one of these intervals. The measure of these intervals is

$$\sum_{q \geq Q} q \cdot \frac{2\psi(q)}{q} < \varepsilon,$$

which proves the theorem. •

For example, we may take $\psi(q) = q^{-1}(\log q)^{-1-\varepsilon}$. Thus almost all numbers cannot be approximated by an order $2+\varepsilon$. A more subtile characterization of real numbers with respect to the order of approximation seems to be a rather difficult task. We will return to this problem in a later section.

We shall briefly give an interesting application of Dirichlet's approximation theorem to dense sequences. As we have seen above for any given real α we can find integers q for which $q\alpha$ differs from an integer by as little as we please. Kronecker's celebrated approximation theorem from 1891 considers the inhomogeneous case.

Theorem 6 *If α is irrational, $\eta \in \mathbb{R}$ is arbitrary, then for any $N \in \mathbb{N}$ there exist $Q \in \mathbb{N}$ with $Q > N$ and $P \in \mathbb{Z}$ such that*

$$|Q\alpha - P - \eta| < \frac{3}{Q}.$$

Proof. By Theorem 4 there are integers $q > 2N$ and p such that

$$|q\alpha - p| < \frac{1}{q}.$$

Suppose that m is the integer, or one of the two integers, for which

$$|m\eta - q| \leq \frac{1}{2}.$$

In view of Theorem 3 and in addition (7) we can write $m = px - qy$, where x and y are integers and $|x| \leq \frac{1}{2}q$. Since

$$q(x\alpha - y - \eta) = x(q\alpha - p) - (q\eta - m),$$

we find

$$|q(x\alpha - y - \eta)| < \frac{1}{2}q \cdot \frac{1}{q} + \frac{1}{2} = 1.$$

Setting $Q = q + x$ and $P = p + y$ yields

$$N < \frac{1}{2}q \leq Q \leq \frac{3}{2}q,$$

and thus

$$|Q\alpha - P - \eta| \leq |x\alpha - y - \eta| + |q\alpha - p| < \frac{2}{q} \leq \frac{3}{N}.$$

This proves the theorem. •

In particular, Kronecker's approximation theorem has interesting consequences on dense sequences:

Exercise 6 *Show that the sequence* $(\{n\alpha\})$ *lies dense in* $[0,1]$ *if and only if* α *is irrational.*

Much more can be said about such sequences. In fact, one can show that *if and only if* α *is irrational, the sequence* $(n\alpha)$ *is* **uniformly distributed modulo** 1, i.e., the *correct* proportion of points $\{n\alpha\}$ lies in an arbitrary subinterval (a,b) of $(0,1)$. On the other side, it is not difficult to show that for example the fractional parts of $\log n$ *prefer small* values where, as usual in number theory, $\log n$ is the logarithm to base e. But it is yet unproved that the sequence $\exp(n)$ is uniformly distributed modulo 1. For a rigorous definition of uniform distribution, the current knowledge in this field and its various applications we refer the reader to [23].

Another interesting application of Kronecker's approximation theorem is the problem of the reflected light ray in plane geometry due to König and Szücs. The sides of a square are reflecting mirrors. A ray of light leaves a point inside the square and is reflected in the mirrors. One can show that *its path is closed and periodic if and only if the angle between a side of the square and the initial direction of the ray has a rational tangent; otherwise the ray of light passes arbitrarily near to every point of the square.* For the proof of this entertaining billiards see [21], §23.3.

In the following sections we will meet two different approaches for more detailed studies concerning diophantine approximations.

3 The Farey sequence

Farey fractions were introduced by Haros in 1802 and (independently) by Farey in 1816. However, Cauchy was the first who studied them systematically.

For any positive integer n the **Farey sequence** $\mathcal{F}_n$ **of order** n is the ordered list of all reduced fractions in the unit interval having denominators $\leq n$:

$$\mathcal{F}_n = \left\{ \frac{a}{b} \in \mathbb{Q} \,:\, 0 \leq a \leq b \leq n \quad \text{with} \quad (a,b) = 1 \right\}.$$

For example,

$$\mathcal{F}_1 = \left\{ \frac{\mathbf{0}}{\mathbf{1}}, \frac{\mathbf{1}}{\mathbf{1}} \right\} \subset \mathcal{F}_2 = \left\{ \frac{0}{1}, \frac{\mathbf{1}}{\mathbf{2}}, \frac{1}{1} \right\} \subset \mathcal{F}_3 = \left\{ \frac{0}{1}, \frac{\mathbf{1}}{\mathbf{3}}, \frac{1}{2}, \frac{\mathbf{2}}{\mathbf{3}}, \frac{1}{1} \right\} \subset \ldots . \tag{13}$$

Clearly, $\mathcal{F}_n \subset \mathcal{F}_{n+1}$. Further, each rational in the unit interval occurs sooner or later in the Farey sequence. Hence, the Farey sequence is building $\mathbb{Q}$ modulo $\mathbb{Z}$. In particular, this construction proves that $\mathbb{Q}$ is countable. We can be a bit more precise to get a feeling for the size of $\mathcal{F}_n$. It is easily seen that the number of Farey fractions in $\mathcal{F}_n$ is related to Euler's totient $\varphi(b)$ which counts the number of positive coprime integers $a \leq b$. With some knowledge on arithmetical functions one can show the asymptotic formula

$$\sharp\mathcal{F}_n = 1 + \sum_{b \leq n} \varphi(b) = \frac{3}{\pi^2}n^2 + O(n \log n),$$

where $f(x) = O(g(x))$ with a positive function $g(x)$ denotes that

$$\limsup_{x \to \infty} \frac{|f(x)|}{g(x)} \quad \text{is bounded.}$$

Two consecutive elements in $\mathcal{F}_n$ are called **neighbours**. One fundamental property of Farey fractions is the spacing of neighbours in the unit interval.

Theorem 7 *For any neighbours* $\frac{a}{b} < \frac{c}{d}$ *in* $\mathcal{F}_n$*,*

$$bc - ad = 1.$$

In particular, under the assumptions of the theorem

$$\frac{c}{d} - \frac{a}{b} = \frac{bc - ad}{bd} = \frac{1}{bd}.$$

This shows that $\mathcal{F}_n$ is not uniformly, but in a *diophantine* sense, optimally distributed, namely with regard to the denominators.

Proof. Consider the diophantine equation

$$bX - aY = 1.$$

Since a and b are coprime, by Theorem 3 and in addition (8) there exists a solution $x, y \in \mathbb{Z}$ with $n - b < y \leq n$. Consequently, also x and y are coprime, and therefore $\frac{x}{y} \in \mathcal{F}_n$. It is easy to compute that

$$\frac{x}{y} = \frac{a}{b} + \frac{1}{by} > \frac{a}{b}. \tag{14}$$

Now suppose that $\frac{c}{d} < \frac{x}{y}$, then

$$\frac{x}{y} - \frac{c}{d} = \frac{dx - cy}{dy} \geq \frac{1}{dy}.$$

Further,

$$\frac{c}{d} - \frac{a}{b} = \frac{bc - ad}{bd} \geq \frac{1}{bd}.$$

These estimates imply

$$\frac{x}{y} - \frac{a}{b} = \frac{x}{y} \underbrace{-\frac{c}{d} + \frac{c}{d}}_{=0} - \frac{a}{b} \geq \frac{1}{bd} + \frac{1}{dy} = \frac{y + b}{bdy} > \frac{n}{bdy}.$$

In view of (14) it follows that $n < d$, contradicting $\frac{c}{d} \in \mathcal{F}_n$. Thus we have $\frac{c}{d} = \frac{x}{y}$ which proves the theorem. •

Note that the proof gives a rule for the computation of the successor of a Farey fraction $\frac{a}{b}$ in $\mathcal{F}_n$. This term is also related to the former *right* neighbour of $\frac{a}{b}$. We define the **mediant** of $\frac{a}{b}, \frac{c}{d} \in \mathcal{F}_n$ by $\frac{a+c}{b+d}$ (like adding two fractions the wrong way). If $\frac{a}{b} < \frac{c}{d}$, then it is easily seen that

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

The mediant is a mean value, but different to other more common mean values like the arithmetic or the geometric mean value, the mediant is not monotonic: for example,

$$\frac{1}{6} < \frac{1}{3} \quad \text{and} \quad \frac{23}{36} < \frac{2}{3}, \quad \text{but} \quad \frac{1+23}{6+36} = \frac{4}{7} > \frac{1}{2} = \frac{1+2}{3+3}.$$

This phenomenon is well-known in statistics where it is called **Simpson's paradox** (though it is not paradox).

Exercise 7 *Show that for any given pair $\frac{A}{B} < \frac{C}{D}$ there exist infinitely many pairs $\frac{a}{b} < \frac{A}{B}, \frac{c}{d} < \frac{C}{D}$ such that*

$$\frac{a+c}{b+d} > \frac{A+B}{C+D}.$$

For this and more see [53].

The importance of the notion of mediants becomes clear by investigating (13). The process of taking mediants builds the Farey sequence out of $\frac{0}{1}$ and $\frac{1}{1}$.

Theorem 8 *The fractions which belong to $\mathcal{F}_n$ but not to $\mathcal{F}_{n-1}$ are mediants of elements of $\mathcal{F}_n$.*

In particular, the property for two consecutive Farey fractions $\frac{a}{b}, \frac{c}{d}$ of being neighbours gets lost in $\mathcal{F}_n$ if $n \geq b + c$.

Proof. With regard to Theorem 7 we have for consecutive Farey fractions

$$\frac{a}{b} < \frac{x}{y} < \frac{c}{d}$$

in $\mathcal{F}_n$ the identities

$$bx - ay = 1 \qquad \text{and} \qquad cy - dx = 1.$$

Multiplying this with c and a, resp. d and b, yields

$$x(bc - ad) = a + c \qquad \text{and} \qquad y(bc - ad) = b + d,$$

which implies the second assertion. •

The Farey sequence is related to an amazing geometry, discovered by Ford in the 1940s. We may embed the unit interval $[0,1]$ into the complex plane $\mathbb{C}$ and define for each $\frac{a}{b} \in \mathcal{F}_n$ the so-called **Ford circle**

$$\mathcal{C}\left(\frac{a}{b}\right) = \left\{ z \in \mathbb{C} \, : \, \left| z - \left(\frac{a}{b} + \frac{i}{2b^2} \right) \right| = \frac{1}{2b^2} \right\},$$

where as usual $i := \sqrt{-1}$. Ford circles provide an interesting sphere packing:

Exercise 8 *Prove that two distinct Ford circles i) have an empty intersection of the interiors, and ii) are tangent if and only if they are associated to neighbours in the Farey sequence.*

We return to our problem to find *explicitly* good approximations to a given α in the unit interval. If we want to get a rational approximation with denominator $\leq n$, then we only have to look for the Farey fraction $\frac{a}{b} \in \mathcal{F}_n$ with minimal distance to α. Further approximations can be found by taking the mediants. It is obvious how one can use Ford circles to localize visually these Farey fractions.

Exercise 9 *The* **golden ratio** *is given by $g = \frac{1}{2}(\sqrt{5} - 1)$. Find all best approximations $\frac{p}{q}$ to g with a denominator $q \leq 100$. Do you see some hidden law? [Compare your observations with the numbers for which the Euclidean algorithm is extremal; see Exercise 2.] Compute the quantities $q|qg - p|$. Do the same for $\alpha = \sqrt{2}$.*

The Farey dissection of the continuum is a very usefool tool in the theory of diophantine approximations. It provides not only an approach to explicit approximations but also an improvement of Dirichlet's approximation theorem, found by Hurwitz in 1891.

Theorem 9 *If α is irrational, then there exist infinitely many rational numbers $\frac{p}{q}$ such that*

$$\left| \alpha - \frac{p}{q} \right| < \frac{1}{\sqrt{5}q^2};$$

the constant $\sqrt{5}$ is best possible.

Proof. Suppose that $\frac{a}{b} < \frac{c}{d}$ are those neighbours in $\mathcal{F}_n$ for which

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

Without loss of generality we assume that $\alpha > \frac{a+c}{b+d}$ (since we may replace α by $1 - \alpha$ otherwise). If now

$$\alpha - \frac{a}{b} \geq \frac{1}{\sqrt{5}b^2}, \quad \alpha - \frac{a+c}{b+d} \geq \frac{1}{\sqrt{5}(b+d)^2} \quad \text{and} \quad \frac{c}{d} - \alpha \geq \frac{1}{\sqrt{5}d^2}, \tag{15}$$

then we obtain by adding these inequalities

$$\frac{c}{d}-\frac{a}{b}\geq\frac{1}{\sqrt{5}}\left(\frac{1}{b^2}+\frac{1}{d^2}\right)=\frac{1}{\sqrt{5}}\frac{b^2+d^2}{b^2d^2}$$

and

$$\frac{c}{d}-\frac{a+c}{b+d}\geq\frac{1}{\sqrt{5}}\left(\frac{1}{(b+d)^2}+\frac{1}{d^2}\right)=\frac{1}{\sqrt{5}}\frac{(b+d)^2+d^2}{(b+d)^2d^2}.$$

On the other side, Theorem 7 and 8 give upper bounds for these quantities, which lead to the inequalities

$$\sqrt{5}bd\geq b^2+d^2 \qquad \text{and} \qquad \sqrt{5}(b+d)d\geq(b+d)^2+d^2.$$

Adding both inequalities gives

$$\sqrt{5}(d^2+2bd)\geq 3d^2+2bd+2b^2,$$

resp.

$$0\geq 4b^2-4(\sqrt{5}-1)bd+(5-2\sqrt{5}+1)d^2=(2b-(\sqrt{5}-1)d)^2,$$

which is impossible since $b,d\in\mathbb{Z}$ but $\sqrt{5}\not\in\mathbb{Q}$. Thus, with view to (15) at least one of the inequalities

$$\left|\alpha-\frac{a}{b}\right|<\frac{1}{\sqrt{5}b^2},\quad \left|\alpha-\frac{a+c}{b+d}\right|<\frac{1}{\sqrt{5}(b+d)^2}\quad\text{and}\quad\left|\alpha-\frac{c}{d}\right|<\frac{1}{\sqrt{5}d^2}$$

holds. Since α is irrational this argument yields the existence of infinitely many such rational approximations by sending $n\to\infty$. The first assertion of the theorem is proved.

For the second one recall the definition of the golden ratio $g=\frac{1}{2}(\sqrt{5}-1)$. We consider the polynomial $P(X):=X^2+X-1$. With $G:=\frac{1}{2}(\sqrt{5}+1)$ we have $P(X)=(X-g)(X+G)$. Now assume that

$$\left|g-\frac{p}{q}\right|<\frac{1}{Cq^2}$$

with $p\in\mathbb{Z},q\in\mathbb{N}$. Then

$$\left|P\left(\frac{p}{q}\right)\right|=\left|g-\frac{p}{q}\right|\cdot\left|G+\frac{p}{q}\right|=\left|g-\frac{p}{q}\right|\cdot\left|G\underbrace{+g-g}_{=0}-\frac{p}{q}\right|<\frac{\sqrt{5}}{Cq^2}+\frac{1}{C^2q^4}.$$

On the other side,

$$\left|P\left(\frac{p}{q}\right)\right|=\frac{|p^2+pq-q^2|}{q^2}\geq\frac{1}{q^2},$$

which implies $C\leq\sqrt{5}$. This proves the second assertion. •

By the proof it seems that the restriction on C depends mainly on the fact that we investigated approximations to g. We call a real number α **quadratic irrational** if it is the root of an irreducible quadratic polynomial with integral coefficients, the **minimal polynomial** of α. Obviously, any such α is of the form

$$\alpha = \frac{a + b\sqrt{d}}{c} \qquad \text{with} \quad a, b \in \mathbb{Z},\ c, d \in \mathbb{N}, \tag{16}$$

where d is **squarefree**, i.e., all prime divisors of d have multiplicity one; in particular, a quadratic irrational is irrational.

Exercise 10 *Suppose that α is the real quadratic irrational with minimal polynomial $aX^2 + bX + c$. Show that for $C > \sqrt{b^2 - 4ac}$ the inequality*

$$\left| \alpha - \frac{p}{q} \right| < \frac{1}{Cq^2}$$

has only finitely many solutions $p, q \in \mathbb{Z}$. Compare this with Exercise 9. Try to formulate a conjecture for quadratic irrationals.

Another interesting topic in context of the Farey sequence is the **Riemann hypothesis** which claims that the prime numbers are as uniformly distributed as possible, namely

$$\pi(x) = \int_2^x \frac{\mathrm{d}u}{\log u} + O\left(x^{\frac{1}{2}+\varepsilon}\right) \qquad \text{for any} \quad \varepsilon > 0, \tag{17}$$

where $\pi(x)$ counts the number of primes $\leq x$. This is problem number 8 of Hilbert's famous list of 23 problems which he presented at the International Congress of Mathematicians in Paris in 1900. The Riemann hypothesis is yet unsolved (actually it is one of the seven Millenium problems). Surprisingly, is related to the deviation of the Farey sequence from being uniformly distributed. Denote the j-th element of $\mathcal{F}_n$ by f_j (ordered by their size). Then, by a result of Franel, *Riemann's hypothesis is true if and only if*

$$\sum_{j=1}^{\sharp\mathcal{F}_n} \left| f_j - \frac{j-1}{\sharp\mathcal{F}_n} \right| = O\left(n^{\frac{1}{2}+\varepsilon}\right)$$

(see [24] for the details).

Further applications of Farey fractions are *error-free computing* strategies (details can be found in [46], §5.12) and the dissection of the complex unit circle with regard to the *circle method* (see [24]).

4 Continued fractions

The powerful tool of continued fractions was first systematically studied by the dutch astronomer Huygens in the 17-th century, motivated by technical problems while constructing a model of our solar system. The most detailed textbook on this topic is the classic [38].

Recall the intimate relation between the approximations to a given rational $\frac{a}{b}$ and the solutions of linear diophantine equations (7), obtained by the Euclidean algorithm. We may rewrite the Euclidean algorithm (6) by

$$\frac{r_{j-1}}{r_j} = \left[\frac{r_{j-1}}{r_j}\right] + \frac{r_{j+1}}{r_j} \qquad \textsf{with} \quad 0 \le r_{j+1} < r_j. \qquad \textsf{for} \quad j = 0, 1, \ldots, n \tag{18}$$

until $r_n \neq 0$. Setting $a_j = \left[\frac{r_{j-1}}{r_j}\right]$, we obtain

$$\frac{r_{-1}}{r_0} = a_0 + \left(\frac{r_0}{r_1}\right)^{-1} = a_0 + \frac{1}{a_1 + \left(\frac{r_1}{r_2}\right)^{-1}} = \ldots.$$

The first of these identities yields the integral part as an approximation to the rational $\frac{r_1}{r_0}$. Using more and more of these identities, we obtain better and better approximations, and among them we can even find all best approximations. We shall give an example: a solar year has

$$\text{365 days 5 hours 48 minutes and 45,8 seconds} \quad \approx \quad 365 + \frac{\mathbf{419}}{\mathbf{1730}} \text{ days.}$$

Unfortunately, this is not an integer, so *how to create a good calendar?* With view to (18) we find

$$\frac{\mathbf{1730}}{\mathbf{419}} = \mathbf{4} + \frac{54}{419},$$

resp.

$$365 + \frac{\mathbf{419}}{\mathbf{1730}} = 365 + \left(\frac{\mathbf{1730}}{\mathbf{419}}\right)^{-1} \approx 365 + \frac{1}{\mathbf{4}},$$

which is nothing else than Caesar's calendar, i.e., a leap year each fourth year. By the full Euclidean algorithm we get

$$365 + \frac{\mathbf{419}}{\mathbf{1730}} = 365 + \cfrac{1}{4 + \cfrac{1}{7 + \cfrac{1}{1 + \cfrac{1}{3 + \cfrac{1}{6 + \cfrac{1}{2}}}}}}.$$

Using this without the last fraction $\frac{1}{2}$ gives

$$365 + \frac{\mathbf{419}}{\mathbf{1730}} \approx 365 + \frac{194}{801};$$

this our present calendar (six leap years are deleted in 800 years), the Gregorian calender, introduced by pope Gregor XIII in 1582.

Exercise 11 *The lunar month has* 29.53 *days; approximate the solar year by* 365.24. *Recover the Metonic cycle of seven leap month every* 19 *years from a good approximation to* $\frac{365.24}{29.53}$ *by use of the Euclidean algorithm. This is used in the jewish calendar.*

We call

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \quad \cdots \quad + \cfrac{1}{a_{n-1} + \cfrac{1}{a_N}}}},$$

where $a_0 \in \mathbb{Z}$ and $a_n \in \mathbb{N}$ for $1 \le n < N$ and $a_N \ge 1$, a **finite simple continued fraction**. Here means *simple* that all *numerators* are equal to one; it is possible to allow other values but in what follows we will, more or less, deal only with simple continued fractions; therefore we may omit the word *simple* in the sequel. The a_n are called **partial denominators**. For brevity we denote the continued fraction above $[a_0, a_1, a_2, \dots, a_N]$.

First, we shall consider $[a_0, \dots, a_N]$ as a function in the variables $a_0, \dots a_N$. We find by induction on n

$$\begin{aligned} [a_0, a_1, \dots, a_n] &= \left[a_0, a_1, \dots, a_{n-1} + \frac{1}{a_n}\right] \\ &= a_0 + \frac{1}{[a_1, \dots, a_n]} = [a_0, [a_1, \dots, a_n]]. \end{aligned} \tag{19}$$

For $n \le N$ we call $[a_0, a_1, \dots, a_n]$ the n**-th convergent** to $[a_0, a_1, \dots, a_N]$ and define

$$\begin{aligned} p_{-1} = 1 \quad , \quad p_0 = a_0 \qquad &\text{and} \qquad p_n = a_n p_{n-1} + p_{n-2}, \\ q_{-1} = 0 \quad , \quad q_0 = 1 \qquad &\text{and} \qquad q_n = a_n q_{n-1} + q_{n-2}. \end{aligned} \tag{20}$$

The computation of the convergents is easily ruled by means of

Theorem 10 *The functions* p_n, q_n *satisfy*

(i) $\frac{p_n}{q_n} = [a_0, a_1, \dots, a_n]$,

(ii) $p_n q_{n-1} - p_{n-1} q_n = (-1)^n$,

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

Proof We prove the first assertion by induction on n. The cases $n = 0$ is obvious. The case $n = 1$ is easily computed by

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

Now assume that the formula in question holds for n. Then, in view of (19) and the recursion formulae for the p_n, q_n

$$\begin{aligned}[a_0, a_1, \ldots, a_n, a_{n+1}] &= \left[a_0, a_1, \ldots, a_n + \frac{1}{a_{n+1}}\right] \\ &= \frac{\left(a_n + \frac{1}{a_{n+1}}\right) p_{n-1} + p_{n-2}}{\left(a_n + \frac{1}{a_{n+1}}\right) q_{n-1} + q_{n-2}} = \frac{a_{n+1}p_n + p_{n-1}}{a_{n+1}q_n + q_{n-1}} = \frac{p_{n+1}}{q_{n+1}},\end{aligned}$$

which proves *(i)*. In view of this we have

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

Repeating this argument with $n-1, n-2, \ldots, 2$ proves *(ii)*. Consequently, *p_n and q_n are coprime.* Further,

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

and the last assertion follows from the second one. •

The continuous fraction expansion is not unique since

$$[a_0, a_1, a_2, \ldots, a_N] = [a_0, a_1, a_2, \ldots, a_N - 1, 1],$$

but *nearly* unique. With regard to our previous observations it is easily seen that *any rational number has a representation as a finite continued fraction $[a_0, a_1, a_2, \ldots, a_N]$, which is unique if $1 < a_N \in \mathbb{N}$.* By Theorem 10 we get a better understanding between the solutions of linear diophantine equations (7) and best approximations:

Exercise 12 *Calculate the continued fraction for $\frac{15}{38}$ and compare its convergents with the approximations found in Section 1. Give a further proof of Exercise 3 by use of Theorem 10.*

We can rewrite the algorithm for computing the continued fraction of a rational $\alpha =: \alpha_0$ by the iteration

$$\alpha_n = [\alpha_n] + \frac{1}{\alpha_{n+1}} \qquad \text{for} \quad n = 0, 1, \ldots\ ,$$

and setting $a_n = [\alpha_n]$. Obviously, if α is rational the iteration stops after finitely many steps. Otherwise, if α is irrational, then the iteration does not stop and we get by this procedure

$$\alpha = [a_0, a_1, a_2, \ldots].$$

$[a_0, a_1, a_2, \ldots]$ is an **infinite continued fraction** but the first thing we have to ask is *whether the underlying infinite process is convergent?* Theorem 10 remains valid and with regard to its second assertion

$$\alpha - \frac{p_n}{q_n} = \frac{\alpha_{n+1}p_n + p_{n-1}}{\alpha_{n+1}q_n + q_{n-1}} - \frac{p_n}{q_n} = \frac{(-1)^n}{q_n(\alpha_{n+1}q_n + q_{n-1})}. \tag{21}$$

Since the q_n are strictly increasing for $n \geq 2$, the sequence of convergents is alternating

$$\frac{p_0}{q_0} < \frac{p_2}{q_2} < \ldots < \alpha < \ldots < \frac{p_3}{q_3} < \frac{p_1}{q_1}.$$

Thus, we have

Theorem 11 *If α is irrational, then*

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

in particular,

$$\alpha = \lim_{n\to\infty} \frac{p_n}{q_n} = [a_0, a_1, a_2, \ldots].$$

It is easily shown that the continued fraction expansion of any irrational number is uniquely determined. Hence, conversely, we can introduce the set of real numbers $\mathbb{R}$ via continued fractions!

With view to Theorem 11 it becomes visible what an important role continued fractions play in the theory of diophantine approximations. Firstly, it gives a third proof of Dirichlet's approximation theorem. In view of (21) we have even

Exercise 13 *Prove that among two consecutive convergents to any irrational α there is at least one satisfying*

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

Furthermore, if $\frac{p}{q}$ is a solution of the latter inequality, then $\frac{p}{q}$ is a convergent to α.

One can go further. Actually, among three consecutive convergents to any irrational α there is at least one satisfying the inequality in Hurwitz' theorem. Thus, we obtain *explicitly* as many *good* approximations to a given irrational as we please. For instance, a short computation gives the sequence of convergents to π

$$[3] = 3, \quad [3,7] = \frac{22}{7}, \quad [3,7,15] = \frac{333}{106},$$
$$[3,7,15,1] = \frac{355}{113}, \quad [3,7,15,1,292] = \frac{1\,03993}{33102}, \quad \ldots,$$

which are located as follows

$$3 < \frac{333}{106} < \frac{1\,03993}{33102} < \ldots < \pi < \ldots < \frac{355}{113} < \frac{22}{7};$$

we observe that the sequence of convergents covers the best approximations! This is not a miracle as Lagrange proved in 1770:

Theorem 12 *Let α be any irrational number with convergents $\frac{p_n}{q_n}$. If $n \geq 2$ and p, q are positive integers satisfying $0 < q \leq q_n$ and $\frac{p}{q} \neq \frac{p_n}{q_n}$, then*

$$|q_n\alpha - p_n| < |q\alpha - p|.$$

This so-called **law of best approximations** shows that we cannot do better than approximating an irrational number by its convergents!

Proof. We may suppose that p and q are coprime. Since

$$|q_n\alpha - p_n| < |q_{n-1}\alpha - p_{n-1}|,$$

it is sufficient to prove the assertion under the assumption that $q_{n-1} < q \leq q$; the full statement of the theorem follows then by induction on n. If $q = q_n$, then $p \neq p_n$ and it follows that

$$\left|\frac{p}{q} - \frac{p_n}{q_n}\right| \geq \frac{1}{q_n}.$$

But

$$\left|\alpha - \frac{p_n}{q_n}\right| \leq \frac{1}{q_nq_{n+1}} < \frac{1}{2q_n}$$

by Theorem 11. Thus

$$\left|\alpha - \frac{p_n}{q_n}\right| < \left|\alpha - \frac{p}{q}\right|,$$

which implies the inequality in question after multiplication with $q = q_n$. Now suppose that $q_{n-1} < q < q_n$. The system of linear equations

$$p_nX + p_{n-1}Y = p \qquad \text{and} \qquad q_nX + q_{n-1}Y = q$$

has the unique solution

$$x = \frac{pq_{n-1} - qp_{n-1}}{p_nq_{n-1} - p_{n-1}q_n} = \pm(pq_{n-1} - qp_{n-1}) \ , \quad y = \frac{pq_n - qp_n}{p_nq_{n-1} - p_{n-1}q_n} = \pm(pq_n - qp_n).$$

Hence, x and y are integers, neither is zero. Obviously, x and y have opposite sign. Since $q_n\alpha - p_n$ and $q_{n-1}\alpha - p_{n-1}$ have opposite sign as well, $x(q_n\alpha - p_n)$ and $y(q_{n-1}\alpha - p_{n-1})$ have the same sign. Since

$$q\alpha - p = x(q_n\alpha - p_n) + y(q_{n-1}\alpha - p_{n-1}),$$

we obtain

$$|q\alpha - p| > |q_{n-1}\alpha - p_{n-1}| > |q_n\alpha - p_n|,$$

which was to show. •

5 The irrationality of $\zeta(3)$

The **Riemann zeta-function** is defined by

$$\zeta(s)=\sum_{n=1}^{\infty}\frac{1}{n^s}=\prod_p\left(1-\frac{1}{p^s}\right)^{-1}$$

for $s>1$, where the so-called **Euler product** is taken over all prime numbers. The identity between the series and the product may be regarded as an analytic version of the *unique prime factorization in* $\mathbb{Z}$. This analytic approach was found by Euler in 1737 and gives a first glance on the intimate relation between $\zeta(s)$ and the distribution of prime numbers. For instance, assume that there are only finitely many primes, then the product converges throughout the complex plane, in contradiction to the divergence of the harmonic series (i.e, the singularity of $\zeta(s)$ at $s=1$). Thus, *there are infinitely many prime numbers.* However, for more deeper studies one has to consider $\zeta(s)$ as a function of a complex variable. (see [24] for more details on this amazing link between analysis and number theory).

The values of the zeta-function taken at the integers are of special interest in number theory. Euler showed that

$$\zeta(2n)=(-1)^{n-1}\frac{(2\pi)^{2n}}{2(2n)!}B_{2n}\qquad\text{for}\quad n\in\mathbb{N},$$

where B_k is the k-th **Bernoulli number** defined by

$$\frac{z}{\exp(z)-1}=\sum_{k=0}^{\infty}\frac{z^k}{k!}B_k;$$

note that B_k is rational. However, nearly nothing is known about the values taken at the odd integers (this is related to the **trivial zeros** of $\zeta(s)$ at $s=-2n, n\in\mathbb{N}$). It was a senstaion when Apéry proved in 1978

Theorem 13 $\zeta(3)$ *is irrational.*

We can only give a sketch of the proof; for the exciting story behind and more details on the proof see [39]. The first step is the recursion formula

$$n^3u_n+(n-1)^3u_{n-2}=(34n^3-51n^2+27n-5)u_{n-1}\qquad\text{for}\quad n\geq 2.$$

We define sequences a_n and b_n generated by this recursion and the initial values

$$a_0=0\ ,\quad a_1=6\qquad\text{and}\qquad b_0=1\ ,\quad b_1=5.$$

It turns out that $b_n\in\mathbb{Z}$ and $a_n\in\mathbb{Q}$ with denominators dividing $2[1,2,\ldots,n]^3$, where $[1,2,\ldots,n]$ denotes here the least common multiple of $1,2,\ldots,n$. Multiplying this recursion formula with $u=a$ by b_{n-1} and additionally the same with $u=b$ by a_{n-1}, we get

$$n^3(a_nb_{n-1}-a_{n-1}b_n)=(n-1)^3(a_{n-1}b_{n-2}-a_{n-2}b_{n-1}).$$

Since $a_1b_0 - a_0b_1 = 6$ we obtain

$$a_nb_{n-1} - a_{n-1}b_n = \frac{6}{n^3}. \tag{22}$$

This leads to an example of a non-simple continued fraction

$$\zeta(3) = \cfrac{6}{5 - \cfrac{1^6}{117 - \cfrac{\ddots}{- \cfrac{n^6}{34n^3 + 51n^2 + 27n + 5 - \ldots}}}} \quad ;$$

this differs from the expansions which we investigated before but it is convergent too. One can deduce from (22)

$$\left|\zeta(3) - \frac{a_n}{b_n}\right| = \sum_{m=n+1}^{\infty} \frac{6}{m^3 b_m b_{m-1}} = O\left(\frac{1}{b_n^2}\right).$$

It is easy to compute by the recursion formula that

$$b_n \ll \alpha^n \quad , \quad \text{where} \quad \alpha = (1+\sqrt{2})^4.$$

By the *prime number theorem*, which is a weaker form of (17), it can be shown that

$$[1, 2, \ldots, n] = \prod_{p \leq n} p^{\left[\frac{\log n}{\log p}\right]} \leq \prod_{p \leq n} n = n^{\pi(n)} \ll e^n.$$

Now setting $p_n = 2a_n[1, 2, \ldots, n]^3$ and $q_n = 2b_n[1, 2, \ldots, n]^3$ (which are both integers by our remark above), we obtain $q_n \ll \alpha^n e^{3n}$ and

$$|q_n\zeta(3) - p_n| = O\left(q_n^{-\delta}\right) \qquad \text{with} \quad \delta = \frac{\log \alpha - 3}{\log \alpha + 3} = 0.08052\ldots.$$

This proves the theorem in view of Exercise 5.

Apéry's discovery started intensive research on this topic with the aim to obtain similar results for other values of $\zeta(s)$. Beukers found a different proof for $\zeta(3) \notin \mathbb{Q}$ by using Legendre polynomials, i.e., for $n \in \mathbb{N}$ the n-th derivative of the polynomial defined in (2). However, it is still open whether $\zeta(5)$ or any other value taken at positive odd integers is irrational. Recently, Rioval proved that *there are infinitely many irrational numbers in the set of odd zeta-values* and Zudilin showed that *at least one of the four numbers* $\zeta(5), \zeta(7), \zeta(9)$ *and* $\zeta(11)$ *is irrational* (cf. [61]).

6 Quadratic irrationals

Leonrado da Pisa (1170-1250) wrote the influential book *Liber abaci* in which he invented the arabic ciphers and zero to Europe. Furthermore, he introduced the **Fibonacci numbers**

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, \ldots,$$

defined by the recursion

$$F_0 := F_1 := 1 \qquad \text{and} \qquad F_{n+1} = F_n + F_{n-1} \qquad \text{for} \quad n \in \mathbb{N}.$$

In view of (20) it follows immediately that

$$\lim_{n\to\infty} \frac{F_n}{F_{n-1}} = [1, 1, 1, , \ldots], \tag{23}$$

but *what is the value of this limit?* If we denote this limit by x, then $x = 1 + \frac{1}{x}$, which shows that x has to be a root of the quadratic polynomial $X^2 - X - 1 = (X - G)(X + g)$ which we already know from the proof of Hurwitz' theorem. Thus, $G = [1, 1, 1, \ldots]$ and $g = \frac{1}{G} = [0, 1, 1, 1, \ldots]$. This observation gives some information on te Fibonacci sequence. For instance, we obtain by Theorem 10

$$F_n F_{n-2} - F_{n-1}^2 = (-1)^n \qquad \text{for} \quad n \in \mathbb{N}.$$

Another beautiful identity was found by Lucas in 1876 who showed

$$(F_m, F_n) = F_{(m,n)} \qquad \text{for} \quad m, n \in \mathbb{N}.$$

The mathematical journal *Fibonacci Quarterly* is exclusively devoted to the Fibonacci sequence.

Exercise 14 *Prove Bernoulli's formula*

$$F_n = \frac{G^n - (-g)^n}{\sqrt{5}} = \left[\frac{G^n}{\sqrt{5}} + \frac{1}{2}\right] \qquad \textit{for} \quad n \in \mathbb{N},$$

and deduce for the generating series

$$\sum_{m=0}^{\infty} F_m z^m = \frac{z}{1 - z - z^2} \qquad \textit{for} \quad |z| < g.$$

As a second example we shall compute the continued fraction expansion for $\sqrt{2}$. We find

$$\sqrt{2} = 1 + \left(\frac{1}{\sqrt{2} - 1}\right)^{-1} = 1 + (\sqrt{2} + 1)^{-1} = 1 + \frac{1}{2 + (\sqrt{2} - 1)} = [1, 2, 2, 2, \ldots].$$

The paper size DINA4 has a very useful self-similarity property. It has approximately the same proportion when folded in half. It is easy to show that if it would have the same proportion, then this proportion would equal $\sqrt{2}$. By definition, A4 paper has length 29.7 and width 21 centimeters. The proportion is a convergent to $\sqrt{2}$:

$$\frac{29.7}{21} = \frac{99}{70} = [1, 2, 2, 2, 2, 2] \approx \sqrt{2} = [1, 2, 2, 2, \ldots].$$

Both examples of quadratic irrationals from above have some pattern in common: their continued fraction expansion is eventually periodic! We say that the continued fraction $[a_0, a_1, \ldots]$ is **periodic** if there exists an integer k with $a_{n+k} = a_k$ for all sufficiently large n. We write

$$[a_0, a_1, \ldots, a_r, \overline{a_{r+1}, \ldots, a_{r+k}}] = [a_0, a_1, \ldots, a_r, a_{r+1}, \ldots, a_{r+k}, a_{r+1}, \ldots, a_{r+k}, \ldots].$$

Exercise 15 *i) Show for $n \in \mathbb{N}$*

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

ii) Find for each $n \in \mathbb{N}$ the continued fraction expansion for the positive root of the polynomial

$$(6n^2+1)X^2 + 3n(8n^2+1)X - (12n^2+1).$$

Hint: start with the case $n = 1, 2$ and search for a pattern!

The pattern of periodicity is overwhelming.

Theorem 14 *A number $\alpha \in \mathbb{R}\backslash\mathbb{Q}$ is quadratic irrational if and only if its continued fraction expansion is eventually periodic.*

This is Lagrange's theorem; its characterization of quadratic irrationals should be compared with the irrationality criterion for real numbers which says that *a real number is rational if and only if its decimal fraction expansion is eventually periodic.* Here is the very nice

Proof. First, assume that $\alpha = [\overline{a_0, a_1, \ldots, a_{k-1}}]$. Then

$$\alpha = \frac{\alpha p_{k-1} + p_{k-2}}{\alpha q_{k-1} + q_{k-2}},$$

resp.

$$q_{k-1}\alpha^2 + (q_{k-2} - p_{k-1})\alpha - p_{k-2} = 0.$$

Since α is irrational the polynomial $q_{k-1}X^2 + (q_{k-2} - p_{k-1})X - p_{k-2}$ is irreducible, and thus its root α is quadratic irrational. Now suppose that

$$\begin{aligned} \alpha &= [a_0, a_1, \ldots, a_r, \overline{a_{r+1}, \ldots, a_{r+k}}] \\ &= [a_0, a_1, \ldots, a_r, \beta] \qquad \text{with} \quad \beta = [\overline{a_{r+1}, \ldots, a_{r+k}}]. \end{aligned}$$

We have already proved that β is quadratic irrational. Hence

$$\alpha = \frac{\beta p_r + p_{r-1}}{\beta q_r + q_{r-1}}$$

is quadratic irrational as well (since $\mathbb{Q}(\beta)$ is a field).

Conversely, assume that $\alpha = [a_0, a_1, \ldots, a_{n-1}, \alpha_n]$ is quadratic irrational. Then there exist $a, b, c \in \mathbb{Z}$ such that $a\alpha^2 + b\alpha + c = 0$. Substituting

$$\alpha = \frac{\alpha_n p_{n-1} + p_{n-2}}{\alpha_n q_{n-1} + q_{n-2}}$$

we obtain

$$A_n\alpha_n^2 + B_n\alpha_n + C_n = 0, \tag{24}$$

where

$$\begin{aligned}
A_n &= ap_{n-1}^2 + bp_{n-1}q_{n-1} + cq_{n-1}^2,\\
B_n &= 2ap_{n-1}p_{n-2} + b(p_{n-1}q_{n-2} + p_{n-2}q_{n-1}) + 2cq_{n-1}q_{n-2},\\
C_n &= ap_{n-2}^2 + bp_{n-2}q_{n-2} + cq_{n-2}^2.
\end{aligned}$$

Note that $A_n = 0$ would imply that the polynomial in (24) has the root $\frac{p_n}{q_n}$, which contradicts $\alpha \notin \mathbb{Q}$. Thus, $A_nX^2 + B_nX + C_n$ is a quadratic polynomial with root α_n. A short calculation with regard to Theorem 10 shows that its discriminant coincides with the discriminant of (24):

$$B_n^2 - 4A_nC_n = (b^2 - 4ac)(p_{n-1}q_{n-2} - p_{n-2}q_{n-1}) = b^2 - 4ac. \tag{25}$$

In view of Theorem 11

$$p_{n-1} = \alpha q_{n-1} + \frac{\delta_{n-1}}{q_{n-1}} \qquad \text{with} \quad |\delta_{n-1}| < 1.$$

Therefore,

$$\begin{aligned}
A_n &= a\left(\alpha q_{n-1} + \frac{\delta_{n-1}}{q_{n-1}}\right)^2 + bq_{n-1}\left(\alpha q_{n-1} + \frac{\delta_{n-1}}{q_{n-1}}\right) + cq_{n-1}^2\\
&= (a\alpha^2 + b\alpha + c)q_{n-1}^2 + 2a\alpha\delta_{n-1} + a\frac{\delta_{n-1}^2}{q_{n-1}^2} + b\delta_{n-1}\\
&= 2a\alpha\delta_{n-1} + a\frac{\delta_{n-1}^2}{q_{n-1}^2} + b\delta_{n-1}.
\end{aligned}$$

It follows that $|A_n| < 2|a\alpha| + |a| + |b|$. Since $C_n = A_{n-1}$ the same estimate holds for C_n as well. Finally, by (25)

$$B_n^2 \leq 4|A_nC_n| + |b^2 - 4ac| < 4(2|a\alpha| + |a| + |b|)^2 + |b^2 - 4ac|.$$

Since the upper bounds for A_n, B_n and C_n do not depend on n, there are only finitely many different triples A_n, B_n, C_n. Thus we can find a triple A, B, C which occurs at least three times, say as $A_{n_1}, B_{n_1}, C_{n_1}$ and $A_{n_2}, B_{n_2}, C_{n_2}$ and $A_{n_3}, B_{n_3}, C_{n_3}$. Consequently, $\alpha_{n_1}, \alpha_{n_2}$ and α_{n_3} are all roots of the quadratic polynomial $AX^2 + BX + C$, and at least two of them must be equal. If, for example, $\alpha_{n_1} = \alpha_{n_2}$ then $a_{n_1} = a_{n_2}, a_{n_1+1} = a_{n_2+1}, \ldots$. This proves the theorem. •

Exercise 16 *Give an estimate for the length of the period of the continued fraction expansion of a quadratic irrational.*

In particular, the partial denominators of quadratic irrationals are bounded. Nearly nothing is known about the continued fraction expansions of cubic irrationals or even transcendental numbers in general. For such numbers it is conjectured that their sequence of partial denominators is unbounded. This topic is investigated in the *metric theory* of continued fraction. For instance, one can show that *the set of real numbers with bounded partial denominators has measure null.* On the other side one can prove that *if $\psi(n)$ is any positive function, then the inequality*

$$a_n = a_n(\alpha) \geq \psi(n)$$

holds for almost all real α if and only if the series

$$\sum_{n=1}^{\infty} \psi(n)$$

diverges. One of the highlights is the *limit theorem of Gauss-Kusmin*: *if meas$_n(x)$ denotes the measure of the set of $\alpha \in (0,1)$ for which $[a_n, a_{n+1}, \ldots] - a_n < x$, then*

$$\lim_{n\to\infty} \text{meas}_n(x) = \frac{\log(1+x)}{\log 2} \qquad \textit{for any} \quad x \in (0,1).$$

It should be noted that Gauss stated this result in a letter to Laplace but never published a proof; the first published proof was given by Kusmin in 1928. For details to this deep result we refer the interested reader to [27]. It is rather difficult to say anything on sums or products of continued fractions. In 1947 Hall [20] showed that *any real number can be represented as sum of two irrationals whose continued fractions only contain partial denominators ≤ 3*; this beautiful result is related to the geometry of Cantor sets. Another interesting problem is the representation of continued fractions as sums of unit fractions; this is related to the *Erdös-Strauss conjecture* which claims that every positive integer $n > 1$ there exist $a, b, c \in \mathbb{N}$ such that

$$\frac{4}{n} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c}.$$

In [14] it is shown, among others, that *for any given positive integer m there exists a continued fraction of length m being the sum of two unit fractions.*

7 The continued fraction for e

In 1748 Euler found the continued fraction of e but he had no full proof. We shall fill the gaps.

Theorem 15 *For $k \in \mathbb{N}$,*

$$\frac{\exp\left(\frac{2}{k}\right)+1}{\exp\left(\frac{2}{k}\right)-1} = [k, 3k, 5k, 7k, \ldots].$$

Proof. For any non-negative integer n define

$$\begin{aligned}
\alpha_n &= \frac{1}{n!}\int_0^1 x^n(1-x)^n \exp\left(\frac{2x}{k}\right)\,\mathrm{d}x, \\
\beta_n &= \frac{1}{n!}\int_0^1 x^{n+1}(1-x)^n \exp\left(\frac{2x}{k}\right)\,\mathrm{d}x;
\end{aligned}$$

the integrands should be compared with the functions f_n which we used in the proof of the irrationality of e. It is not difficult to compute

$$\alpha_0 = \frac{k}{2}\left(\exp\left(\frac{2}{k}\right) - 1\right), \quad \beta_0 = \frac{k}{2}\exp\left(\frac{2}{k}\right) - \left(\frac{k}{2}\right)^2\left(\exp\left(\frac{2}{k}\right) - 1\right). \tag{26}$$

Exercise 17 *Show for* $n \in \mathbb{N}$ *that*

$$\frac{2}{k}\alpha_n + \alpha_{n-1} = 2\beta_{n-1} \qquad \textit{and} \qquad k(2n+1)\alpha_n = k\beta_{n-1} - 2\beta_n.$$

With regard to these formulae we may eliminate the β_n's:

$$\frac{2}{k}\alpha_{n+1} + k(2n+1)\alpha_n = \frac{k}{2}\left(\frac{2}{k}\alpha_n + \alpha_{n-1}\right) - \alpha_n = \frac{k}{2}\alpha_{n-1},$$

resp.

$$\frac{2\alpha_{n+1}}{k\alpha_n} + (2n+1)k = \frac{k\alpha_{n-1}}{2\alpha_n}. \tag{27}$$

In view of (26) we deduce

$$\alpha_1 = \frac{k}{2}(2\beta_0 - \alpha_0) = \left(\frac{k}{2}\right)^2\left(\exp\left(\frac{2}{k}\right) + 1 - k\left(\exp\left(\frac{2}{k}\right) - 1\right)\right).$$

This leads to

$$\frac{\exp\left(\frac{2}{k}\right) + 1}{\exp\left(\frac{2}{k}\right) - 1} = \frac{\left(\frac{2}{k}\right)^2\alpha_1 + k\left(\exp\left(\frac{2}{k}\right) - 1\right)}{\frac{2}{k}\alpha_0} = k + \frac{2\alpha_1}{k\alpha_0}.$$

It is easily seen that $2\alpha_{n+1} < k\alpha_n$ for $n \in \mathbb{N}$, which yields in view of (27)

$$\frac{\exp\left(\frac{2}{k}\right) + 1}{\exp\left(\frac{2}{k}\right) - 1} = k + \left(\frac{k\alpha_0}{2\alpha_1}\right)^{-1} = \left[k, 3k, \frac{k\alpha_1}{2\alpha_2}\right] = \ldots .$$

This proves the theorem. •

In particular, with regard to Lagrange's theorem we find a slightly improvement of Theorem 1:

Exercise 18 *Prove that* e *is neither rational nor quadratic irrational.*

In what follows we exactly follow Euler in computing the continued fraction expansion for e. Taking $k = 2$ in Theorem 15 we get

$$\frac{e+1}{e-1} = [2, 6, 10, 14, \ldots];$$

denote its n-th convergents by $\frac{p_n}{q_n}$. Further, define the real number E by $E = [A_0, A_1, A_2, \ldots]$, where

$$A_0 = 2,\ A_{3n-2} = A_{3n} = 1 \quad \text{and} \quad A_{3n+1} = 2n,$$

and let $\frac{P_n}{Q_n}$ be the n-th convergent to E.

Exercise 19 *Prove for $n \in \mathbb{N}$ the identities*

$$P_{3n+1} = p_n + q_n \qquad \textit{and} \qquad Q_{3n+1} = p_n - q_n.$$

Hint: for $-3 \leq j \leq 1$ write down P_{3n+j} in terms of P_{3n+j-1} and P_{3n+j-2}, and multiply the resulting five equations by $1, -1, 2, 1$ and 1. Then the identity for P_{3n+1} follows by induction; the formula for Q can be proved analogously.

This implies

$$E = \lim_{n\to\infty} \frac{P_{3n+1}}{Q_{3n+1}} = \lim_{n\to\infty} \frac{\frac{p_n}{q_n}+1}{\frac{p_n}{q_n}-1} = \frac{\frac{e+1}{e-1}+1}{\frac{e+1}{e-1}-1} = e.$$

Thus, we have proved

Theorem 16 $e = [2, 1, 2, 1, 1, 4, 1, \ldots]$.

Euler's approach to continued fractions was rather different. He expanded a certain generalization of the exponential function into a continued fraction (see [29]). This idea can be regarded as the starting point of the so-called *Padé approximations*, which is the theory of approximating transcendental function by rational function.

It is an easy consequence that e is not approximable by an order $\kappa > 2$:

Exercise 20 *Show that*

$$\left| e - \frac{p}{q} \right| > \frac{c}{q^2 \log q}$$

for all $\frac{p}{q} \in \mathbb{Q}$ with $q > 1$ and some positive constant c.

It should be noted that the continued fraction for π shows no pattern so far. Curiously, if we switch to another expansion we may find patterns.

God loves the odd integers. (Leibniz)

From the Leibniz series

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} \pm \ldots,$$

Euler deduced the continued fraction expansion

$$\frac{\pi}{4} = \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{9}{2 + \ddots \cfrac{}{+ \cfrac{(2n+1)^2}{2 + \ddots}}}}} .$$

8 Markoff's spectrum

The constant $\sqrt{5}$ in Hurwitz' theorem depends mainly on best approximations of the golden ratio $\alpha = G$. The same argument as in its proof suggests that in case of $\alpha = \sqrt{2}$ the constant in Hurwitz' theorem can be refined to the larger constant $\sqrt{8}$ (see Exercise 10).

For this purpose we have to introduce some new definitions. We call two real numbers α and β **equivalent** if there are integers a, b, c, d such that

$$\alpha = \frac{a\beta + b}{c\beta + d} \qquad \text{with} \quad ad - bc = \pm 1. \tag{28}$$

Such a mapping $z \mapsto \frac{az+b}{cz+d}$ is called **unimodular transformation** (that is a special **Möbius transformation**). They build the **group of unimodular transformations** (for details on the beautiful theory of modular transformations and their geometry see [30]). Therefore, equivalence of real numbers is symmetric, transitive and reflexive. In view of Theorem 10 each real number α is equivalent to any of its tails α_n, more precisely: $\alpha = [a_0, a_1, \ldots, a_{n-1}, \alpha_n]$ may be regarded as n successive modular transformations. It is easily seen that *any two rational numbers are equivalent* (it suffices to show that each rational is equivalent to zero). The interpretation of the real line via the actions of modular transformations provides a new view on our previous observations on the Farey sequence and on the continued fraction algorithm.

Exercise 21 *Prove that a real number $\alpha = [a_0, a_1, \ldots, a_{n-1}, \beta]$, where $n \geq 2$, has a representation by a modular transformation (28) if and only if $c > d > 0$.*
Hint: use induction on d for the converse implication.

Now we are in the position to compare the continued fraction expansions of equivalent numbers.

Theorem 17 *Two irrational numbers α and β are equivalent if and only if their continued fraction expansions are eventually identical; more precisely: there exist positive integers m, n and a real number $\gamma > 1$ such that*

$$\alpha = [a_0, \ldots, a_n, \gamma] \qquad \textit{and} \qquad \beta = [b_0, \ldots, b_m, \gamma].$$

Proof. By Theorem 10,

$$\alpha = \frac{p_n\gamma + p_{n-1}}{q_n\gamma + q_{n-1}},$$

where $p_nq_{n-1} - p_{n-1}q_n = \pm 1$. Hence, α and γ are equivalent. Similarly, β and γ are equivalent. Since the Möbius transformations form a group, α and β are also equivalent. Conversely, assume that α and β are equivalent. Hence, the identity (28) holds with $c > d > 0$ (by the previous exercise). This together with

$$\beta = [b_0, \ldots, b_m, \gamma] = \frac{p_m\gamma + p_{m-1}}{q_m\gamma + q_{m-1}}$$

implies

$$\alpha = \frac{P\gamma + R}{Q\gamma + S},$$

where

$$P = ap_m + bq_m, \ R = ap_{m-1} + bq_{m-1}, \ Q = cp_m + dq_m, \ S = ap_{m-1} + bq_{m-1},$$

and $PS - QR = \pm 1$ (this is easily seen by considering modular transforms as matrices). In view of Theorem 11

$$p_{m-j} = \alpha q_{m-j} + \frac{\delta_j}{q_{m-j}} \qquad \text{with} \quad |\delta_j| < 1$$

for $j = 0, 1$. Consequently,

$$Q = (c\alpha + d)q_m + \frac{c\delta_0}{q_m} > (c\alpha + d)q_{m-1} + \frac{c\delta_1}{q_{m-1}} = S$$

for sufficiently large m. Obviously, we may assume without loss of generality that $c\alpha + d > 0$, which implies $S > 0$. With view to Exercise 21 it follows that $\alpha = [a_0, \ldots, a_n, \gamma]$, which was to show. •

Now let $\|x\| = \min\{|x - z| \, : \, z \in \mathbb{Z}$. The **Markoff constant** for α is defined by

$$M(\alpha) = \liminf_{q \to \infty} q\|q\alpha\|.$$

In view of Hurwitz' theorem we have $M(\alpha) \leq M(G) = \frac{1}{\sqrt{5}}$ for any real α (in view of (23), which supports our observation $M(g) \approx 0.477\ldots$ from Exercise 9). We need a more explicit representation of the Markoff constant. Taking into account (21) we have

$$\alpha - \frac{p_n}{q_n} = \frac{1}{\lambda_n q_n^2}\ , \qquad \text{where} \quad \lambda_n = (-1)^n \left(\alpha_{n+1} + \frac{q_{n-1}}{q_n} \right)$$

and $\alpha_{n+1} = [a_{n+1}, a_{n+2}, \ldots]$. Since

$$\frac{q_{n-1}}{q_n} = \frac{1}{q_n/q_{n-1}} = \frac{1}{a_n + \frac{q_{n-2}}{q_{n-1}}} = [0, a_n, a_{n-1}, \ldots, a_1],$$

we deduce

$$M(\alpha) = \liminf_{n \to \infty} \frac{1}{\lambda_n} = \liminf_{n \to \infty} \frac{1}{[a_{n+1}, a_{n+2}, \ldots,] + [0, a_n, a_{n-1}, \ldots, a_1, a_0]}.$$

In view of Theorem 17 we obtain

Theorem 18 *If α and β are equivalent, then $M(\alpha) = M(\beta)$.*

In particular, if α is irrational with positive $M(\alpha)$, then there exist infinitely many rationals $\frac{p}{q}$ for which

$$\left|\alpha - \frac{p}{q}\right| \leq \frac{M(\alpha)}{q^2}.$$

This yields the following refinement of Hurwitz' theorem. *Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$. Then $M(\alpha) = \frac{1}{\sqrt{5}}$ if and only if α is equivalent to G. If α is not equivalent to G, then $M(\alpha) \leq M(\sqrt{2}) = \frac{1}{\sqrt{8}}$*; this can be proved along the lines of the proof of Theorem 9. This shows that all real numbers, which are equivalent to the golden ratio, that are those which have an eventually periodic continued fraction with period $\overline{1}$, are the *worst* approximable irrationals (this should be compared with Exercise 2). This is only the first step as Markoff observed. We could continue this process to find successively smaller Markoff constants by further restrictions on the continued fraction expansions. The set of all Markoff constants is the **Markoff spectrum**. This leads to

- $G = [\overline{1}]$ with $M(G) = \frac{1}{\sqrt{5}} = 0.44721\ldots$,
- $1 + \sqrt{2} = [\overline{2}]$ with $M(1 + \sqrt{2}) = \frac{1}{\sqrt{8}} = 0.35355\ldots$,
- $\frac{9+\sqrt{221}}{10} = [\overline{2, 2, 1, 1}]$ with $M\left(\frac{9+\sqrt{221}}{10}\right) = \frac{5}{\sqrt{221}} = 0.33633\ldots$;

for a longer list see [9]. There is a deep *theorem of Markoff* [32] which states that *the Markoff spectrum above $\frac{1}{3}$ consists exactly of numbers of the form*

$$\frac{z}{\sqrt{9z^2 - 4}},$$

where z is a positive integer such that there exist $x, y \in \mathbb{N}$ with $\max\{x, y\} \leq z$ satisfying the diophantine equation

$$X^2 + Y^2 + Z^2 = 3XYZ.$$

Unfortunately, this is far beyond our scope. Further, it can be shown that *any positive real number less than* **Freiman's number**

$$\frac{153640040533216 - 19623586058\sqrt{462}}{693746111282512} = 0.22085\ldots$$

is in the Markoff spectrum. See [49] for a nice survey on the Markoff spectrum and its interaction with *hyperbolic geometry.*

Exercise 22 *Compute the Markoff constant for $\alpha = \sqrt{n^2 + 2}$ for $n \in \mathbb{N}$.*
Hint: see Exercise 15.

We call a number α **badly approximable** if $M(\alpha) > 0$, i.e. there exists a positive constant c, depending only on α, such that for *all* $\frac{p}{q}$

$$\left|\alpha - \frac{p}{q}\right| > \frac{c}{q^2}$$

holds. The following theorem classifies badly approximable numbers.

Theorem 19 *An irrational α is badly approximable if and only if its partial quotients are bounded.*

For instance, quadratic irrationals are badly approximable, but not e (see Theorem 16 and Exercise 20).

Proof. In view of (21)

$$\frac{1}{(a_{n+1}+2)q_n^2} \leq \left|\alpha - \frac{p_n}{q_n}\right| \leq \frac{1}{a_{n+1}q_n^2}. \tag{29}$$

By Theorem 12, the law of best approximations, other rationals cannot approximate α better than convergents do. This proves the theorem. •

9 The Pell equation

In a letter to Eratosthenes Archimedes $(287 - 212$ B.C.) posed the so-called **cattle problem** in which he asks for the number of bulls and cows belonging to the Sun god, subject to certain arithmetical restrictions. This problem was already forgotten over the centuries until it was rediscovered by G.E. Lessing in the Wolfenbüttel library in 1773. A nice English version goes as follows:

> *The Sun god's cattle, apply thy care to count their numbers, hast thou wisdom's share. They grazed of old on the Thrinacian floor of Sic'ly's island, herded in to four, colur by colour: one herd white as cream, the next in coats glowing with ebon gleam, brown-skinned the third, and stained with spots the last. Each herds saw bulls in power unsurpassed, in ratios these: count half the ebon-hued, add one third more, then all the brown include; thus, friend, canst thou the white bulls' number tell. The ebon did the brown exceed as well, now by a fourth and fifth part of the stained. To know the spotted - all bulls that remained - reckon again the brown bulls, and unite these with a sixth and seventh of the white. Among the cows, the tale of silver-haired was, when with bulls and cows of black compared, exactly one in three plus one in four. The black cows counted one in four once more, plus now a fifth, of the bespeckled breed when, bulls withal, they wandered out to feed. The speckled cows tallied a fifth and sixth of all the brown-haired, males and females mixed. Lastly, the brown cows numbered half a third and one in seven of the silver herd. Tellst thou unfailingly how many head the Sun possessed, o friend, both bulls*

well-fed and cows of ev'ry colour - no-one will deny thou hast numbers' art and skill, though not yet dost thou rank among the wise. But come! also the foll'wing recognise. Whene'er the Sun God's white bulls joined the black, their multitude would gather in a pack of equal length and breadth, and squarely throng Thrinacia's territory broad and long. But when the brown bulls mingled with the flecked, in rows growing from one would they collect, forming a perfect triangle, with ne'er a diff'rent-coloured bull, and none to spare. friend, canst thou analyse this in thy mind, and of these masses all the measures find, go forth in glory! be assured all deem thy wisdom in this discipline supreme! (cf. [1])

Denoting by w, x, y and z the numbers of the white, black, dappled, and brown bulls, and by $\mathcal{W}, \mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$ the numbers of the white, black, dappled, and brown cows, respectively, one has to solve the system of linear diophantine equations

$$w = \left(\frac{1}{2} + \frac{1}{3}\right) x + z \ , \quad x = \left(\frac{1}{4} + \frac{1}{5}\right) y + z \ , \quad y = \left(\frac{1}{6} + \frac{1}{7}\right) w + z, \tag{30}$$

and

$$\begin{aligned} \mathcal{W} &= \left(\frac{1}{3} + \frac{1}{4}\right)(x + \mathcal{X}) \ , \qquad & \mathcal{X} &= \left(\frac{1}{4} + \frac{1}{5}\right)(y + \mathcal{Y}) \ , \\ \mathcal{Y} &= \left(\frac{1}{5} + \frac{1}{6}\right)(z + \mathcal{Z}) \ , \qquad & \mathcal{Z} &= \left(\frac{1}{6} + \frac{1}{7}\right)(w + \mathcal{W}). \end{aligned} \tag{31}$$

Due to Archimedes everyone who can solve this problem is *merely competent*; but to win the prize for *supreme wisdom* one has to meet the additional condition that $w + x$ has to be a square, and that $y + z$ has to be a **triangular** number, i.e., a number of the form $1 + 2 + \ldots + n = \frac{1}{2}n(n+1)$ (imagine the numbered balls in pool billiard). It is easily seen that the general solution to (30) is given by

$$(w, x, y, z) = m \cdot (2226, 1602, 1580, 891), \qquad \text{where} \quad m \in \mathbb{N}.$$

It turns out that the system (31) is solvable if and only if m is a multiple of 4657. Setting $m = 4657 \cdot M$ we obtain for the general solution of (31)

$$(\mathcal{W}, \mathcal{X}, \mathcal{Y}, \mathcal{Z}) = M \cdot (7206360, 4893246, 3515820, 5439213), \qquad \text{where} \quad \mathcal{M} \in \mathbb{N}.$$

This solves the first part of the cattle problem. To solve also the second part one has to find an $\mathcal{M}$ such that $w+x = 4657 \cdot 3828 \cdot \mathcal{M}$ is a square and $y+z = 4657 \cdot 2471 \cdot \mathcal{M}$ is triangular. By the prime factorization $4657 \cdot 3828 = 2^2 \cdot 3 \cdot 11 \cdot 29 \cdot 4657$ it follows that $w + x$ is a square if and only if $\mathcal{M} = 3 \cdot 11 \cdot 29 \cdot 4657 \cdot Y^2$, where $Y \in \mathbb{N}$. Since $y + z$ is triangular if and only if $8(y + z) + 1$ is a square, one has to solve the quadratic equation

$$X^2 - 410\,286\,423\,278\,424 \cdot Y^2 = 1,$$

where $410\,286\,423\,278\,424 = 2 \cdot 3 \cdot 7 \cdot 11 \cdot 29 \cdot 353 \cdot (2 \cdot 4657)^2$, which does not look too easy. It seems that the ancient Greek were unable to solve this equation; see [31] or [47] for nicely written surveys on the cattle problem, its history and the first prize winners for supreme wisdom.

We shall even solve a more general problem. The **Pell equation** is defined by

$$X^2 - dY^2 = 1, \qquad \text{where} \quad d \in \mathbb{N}. \tag{32}$$

It should be noted that Pell was an English mathematician who lived in the seventeenth century but he had nothing to do with this equation (it was Euler who mistakenly attributed a solution method to Pell which in fact was found by Pell's contemporaries Wallis and Lord Brouncker). We are interested in integral solutions. In some sense, we have to find the set of intersections of a hyperbola with the lattice $\mathbb{Z}^2$. Obviously, $x = 1$ and $y = 0$ is always a solution, *but are there more?* By symmetry it suffices to look for solutions in positive integers. If d is a perfect square, we can factor the left hand side and it turns out that (32) has only finitely many integral solutions. Thus, the case of a perfect square d is boring and we may assume in the sequel that d is not a perfect square, resp. $\sqrt{d} \not\in \mathbb{Q}$.

Our first deeper observation is due to Euler. Assume that x, y is a solution of (32). Then we may factor the left hand side of (32) which leads to

$$(x - y\sqrt{d})(x + y\sqrt{d}) = x^2 - dy^2 = 1,$$

resp.

$$\left|\sqrt{d} - \frac{x}{y}\right| = \frac{1}{y^2(\sqrt{d} + \frac{x}{y})} < \frac{1}{2y^2}.$$

In view of Exercise 13 all solutions of (32) can be found among the convergents to $\sqrt{d}$. For instance, the sequence of convergents $\frac{p_n}{q_n}$ to $\sqrt{2}$ starts with

$$\frac{1}{1}, \frac{\mathbf{3}}{\mathbf{2}}, \frac{7}{5}, \frac{\mathbf{17}}{\mathbf{12}}, \frac{41}{29}, \frac{\mathbf{99}}{\mathbf{70}}, \ldots \quad \to \quad \sqrt{2} = [1, \overline{2}],$$

and in fact we get for $p_n^2 - 2q_n^2$ the values

$$1^2 - 2 \cdot 1^2 = -1, \quad \mathbf{3}^2 - 2 \cdot \mathbf{2}^2 = +1, \quad 7^2 - 2 \cdot 5^2 = -1,$$

$$\mathbf{17}^2 - 2 \cdot \mathbf{12}^2 = +1, \quad 41^2 - 2 \cdot 29^2 = -1, \quad \mathbf{99}^2 - 2 \cdot \mathbf{70}^2 = +1.$$

The regularity is astonishing! It suggests to consider instead of (32) the more general quadratic equation

$$X^2 - dY^2 = \pm 1; \tag{33}$$

according to the sign we speak about the **plus** and the **minus** equation, respectively.

Exercise 23 *Consider the the cases $d = 11, 13$. Write down the convergents $\frac{p_n}{q_n}$ to $\sqrt{d}$ for the first two periods and compute the values $p_n^2 - dq_n^2$. Give a conjecture how the set of solutions of* (33) *looks like.*

It is easily seen that *if d is a multiple of a prime $p \equiv 3 \bmod 4$, then the minus equation is unsolvable* (since squares are congruent $0, 1 \bmod 4$).

From a mathematical point of view the Pell equation is important with respect to its role in *algebraic number theory.* A complex number α is said to be **algebraic over** $\mathbb{Q}$ if there exists a polynomial $P(X)$ with rational coefficients such that $P(\alpha) = 0$; the polynomial with leading coefficient 1 and minimal degree having this property is called **minimal polynomial** of α and we denote it by $P_\alpha(X)$ (this generalizes our concept of quadratic irrationals); the degree of the minimal polynomial is said to be the **degree** of α; for short $d := \deg \alpha = \deg P_\alpha$. The set

$$\mathbb{Q}(\alpha) = \{a_0 + a_1\alpha + \ldots + a_{d-1}\alpha^{d-1} \; : \; a_j \in \mathbb{Q}\}$$

is a finite algebraic extension of the field of rational numbers, the **algebraic number field** associated to α; it has **degree** $d = [\mathbb{Q}(\alpha) : \mathbb{Q}(\alpha)]$ (which may be regarded as the dimension of the rational vector space $\mathbb{Q}(\alpha)$). Note that any number in $\mathbb{Q}(\alpha)$ is algebraic. The zeros $\alpha_1', \ldots, \alpha_d'$ of the minimal polynomial $P_\alpha(X)$ are the **conjugates** of α (i.e. the images of α under the field automorphisms). The product of all conjugates is the **norm** of α:

$$N(\alpha) := \prod_{j=1}^{d} \alpha_j';$$

note that this is, up to a sign, equal to the constant term $P_\alpha(0)$ in the minimal polynomial; the norm provides a *measure for the size* of algebraic numbers. An algebraic number is said to be an **algebraic integer** if its minimal polynomial has integral coefficients. The set of all algebraic integers in a number field form a ring, the so-called **ring of integers**. Unfortunately, and somehow surprisingly, these rings in general do not have a unique prime factorization any longer. For example,

$$2 \cdot 3 = (1 - \sqrt{-5}) \cdot (1 + \sqrt{-5})$$

gives two distinct factorizations of 6 in the ring $\mathbb{Z}[\sqrt{-5}]$ into irreducible factors (one can overcome this problem by introducing *prime ideals* but this is another story). For an understanding of the structure of number fields it is important to study its integers and, in particular, its units. An algebraic integer is called **unit** if the absolute value of its norm is equal to one.

In case of $\deg \alpha = 2$ we call $\mathbb{Q}(\alpha)$ a **quadratic** number field. It is easily seen that there always exists $d \in \mathbb{Z}$, which is not a perfect square, such that $\mathbb{Q}(\alpha) = \mathbb{Q}(\sqrt{d})$. We say that $\mathbb{Q}(\sqrt{d})$ is a **real** or **imaginary** quadratic number field according to $\sqrt{d} \in \mathbb{R}$ or not. One can show that the ring of integers equals $\mathbb{Z}[\sqrt{d}]$ if $d \equiv 2, 3 \bmod 4$, and $\mathbb{Z}[\frac{1+\sqrt{d}}{2}]$ otherwise. For example, the golden ratio $G = \frac{1+\sqrt{5}}{2}$ is an algebraic integer in $\mathbb{Q}(\sqrt{5})$. The conjugates of algebraic integers are of the form $x \pm y\sqrt{d}$, where x, y are rational with denominators ≤ 2. Therefore, the units of a real quadratic number field are given by the related *rational* solutions of the Pell equation

$$x^2 - dy^2 = (x + y\sqrt{d})(x + y\sqrt{d})' = N(x + y\sqrt{d}) = \pm 1,$$

resp. to the *integral* solutions of the slightly more general equation $X^2 - dY^2 = \pm 1, \pm 4$ according to the residue class of d modulo 4. An imaginary quadratic number field has only finitely many units, each of them being a root of unity. For more details see [10] or [21].

Before we return to the Pell equation we need to introduce the notion of reducibility. A quadratic irrational α is called **reduced** if $\alpha > 1$ and $-1 < \alpha' < 0$. The following important result is due to Galois.

Theorem 20 *The continued fraction expansion of a quadratic irrational number α is purely periodic if and only if α is reduced.*

We prove only the implication which we shall use later on.

Proof. Assume that $\alpha = [a_0, a_1, \ldots, a_{n-1}, \alpha_n]$ is reduced. Then, for $n = 0, 1, \ldots$,

$$\alpha_n = a_n + \frac{1}{\alpha_{n+1}} \qquad \text{and} \qquad \alpha'_n = a_n + \frac{1}{\alpha'_{n+1}}.$$

Consequently, $\alpha_n > 1$. If $\alpha'_n < 0$ it follows that

$$-1 < \alpha'_{n+1} = \frac{1}{\alpha'_n - a_n} < 0.$$

Hence, by induction, all α_n are reduced. In particular,

$$0 < -\frac{1}{\alpha'_{n+1}} - a_n < 1 \ , \qquad \text{resp.} \quad a_n = \left[-\frac{1}{\alpha'_{n+1}} \right].$$

Since α is quadratic irrational, by Lagrange's theorem its continued fraction is eventually periodic. Thus there exist $k < l$ for which $\alpha_k = \alpha_l$. It follows that $a_k = a_l$ and that $\alpha'_k = \alpha'_l$. Thus,

$$a_{k-1} = \left[-\frac{1}{\alpha'_k} \right] = \left[-\frac{1}{\alpha'_l} \right] = a_{l-1}.$$

We conclude by induction that the continued fraction of α is purely periodic. •

Exercise 24 *Assume that $\alpha = [\overline{a_0, \ldots, a_n}]$ is a quadratic irrational. Show that*

$$[\overline{a_n, \ldots, a_0}] = -\frac{1}{\alpha'}.$$

Deduce the other implication in Galois' theorem.
Hint: It suffices to show that if $[a, \overline{b_1, \ldots, b_n}]$ is reduced, then $a = b_n$.

Now we can give a detailed description of the continued fraction expansion of $\sqrt{d}$ due to Legendre and Lagrange.

Theorem 21 *If d is not a perfect square, then*

$$\sqrt{d} = [[\sqrt{d}], \overline{a_1, \ldots, a_{N-1}, 2[\sqrt{d}]}].$$

Proof. Obviously, $-1 < [\sqrt{d}] - \sqrt{d} < 0$. Consequently, $\sqrt{d} + [\sqrt{d}] > 1$ is reduced, and hence by Galois' theorem purely periodic:

$$\sqrt{d} + [\sqrt{d}] = [\overline{2[\sqrt{d}], a_1, \ldots, a_{N-1}}] = [2[\sqrt{d}], \overline{a_1, \ldots, a_{N-1}, 2[\sqrt{d}]}].$$

This implies the representation given in the theorem. •

In what follows $N = N(d)$ denotes the minimal length of the periods of the continued fraction expansion of $\sqrt{d}$. We write

$$\sqrt{d} = [[\sqrt{d}], a_1, \ldots, a_n, \alpha_{n+1}],$$

then

$$\alpha_1 = \frac{1}{\sqrt{d} - [\sqrt{d}]} = \frac{\sqrt{d} + [\sqrt{d}]}{d - [\sqrt{d}]^2} =: \frac{P_1 + \sqrt{d}}{Q_1},$$

where P_1, Q_1 are integral and Q_1 divides $d - P_1^2$. Now assume that

$$\alpha_n = \frac{P_n + \sqrt{d}}{Q_n}, \qquad \text{where} \quad P_n, Q_n \in \mathbb{Z},\ Q_n | (d - P_n^2). \tag{34}$$

Then it follows that

$$\alpha_{n+1} = \frac{1}{\alpha_n - a_n} = \frac{Q_n}{P_n - a_n Q_n + \sqrt{d}} = \frac{Q_n(P_n - a_n Q_n - \sqrt{d})}{(P_n - a_n Q_n)^2 - d} =: \frac{P_{n+1} + \sqrt{d}}{Q_{n+1}},$$

where $P_{n+1} = a_n Q_n - P_n$ and

$$Q_{n+1} = \frac{d - (P_n - a_n Q_n)^2}{Q_n} = \underbrace{\frac{d - P_n^2}{Q_n}}_{\in \mathbb{Z}} + 2a_n P_n - a_n^2 Q_n$$

are integral. Since

$$Q_n = \frac{d - (P_n - a_n Q_n)^2}{Q_{n+1}} = \frac{d - P_{n+1}^2}{Q_{n+1}},$$

it follows that Q_{n+1} divides $d - P_{n+1}^2$. Hence, by induction on n we see that each α_n has a representation of the form (34). Consequently,

$$\sqrt{d} = \frac{\alpha_n p_{n-1} + p_{n-2}}{\alpha_n q_{n-1} + q_{n-2}} = \frac{(P_n + \sqrt{d})p_{n-1} + p_{n-2}Q_n}{(P_n + \sqrt{d})q_{n-1} + q_{n-2}Q_n},$$

resp.

$$\sqrt{d}((P_n + \sqrt{d}Q_n)q_{n-1} + q_{n-2}Q_n) = (P_n + \sqrt{d}Q_n)p_{n-1} + p_{n-2}Q_n.$$

Since $\sqrt{d} \notin \mathbb{Q}$, splitting the latter identity into its rational and its irrational parts yields

$$dq_{n-1} = P_n p_{n-1} + p_{n-2} Q_n \qquad \text{and} \qquad p_{n-1} = P_n q_{n-1} + q_{n-2} Q_n.$$

Multiplying the first one by q_{n-1} and the second one by p_{n-1}, subtraction of both equations gives with regard to Theorem 10

$$p_{n-1}^2 - dq_{n-1}^2 = Q_n(p_{n-1}q_{n-2} - p_{n-2}q_{n-1}) = (-1)^n Q_n. \tag{35}$$

This should be compared with Exercise 23. In particular, if n is multiple of the minimal period N, then

$$\frac{P_{kN} + \sqrt{d}}{Q_{kN}} = \alpha_{kN} = [0, \overline{a_1, \ldots, a_{N-1}}] = \sqrt{d} - [\sqrt{d}],$$

by (34). Consequently, $Q_{kN} = 1$. Thus, the plus-equation has infinitely many solutions in positive integers x_k, y_k given by

$$(x_k, y_k) = \begin{cases} (p_{kN-1}, q_{kN-1}) & \text{if} \quad N \equiv 0 \bmod 2, \\ (p_{2kN-1}, q_{2kN-1}) & \text{if} \quad N \equiv 1 \bmod 2. \end{cases} \tag{36}$$

The minus-equation has infinitely many solutions $x_k = p_{(2k-1)N-1}, y_k = q_{(2k-1)N-1}$ if N is odd; actually, it can be shown that *the minus equation is solvable if and only if the minimal period N is odd*; for another criterion see also Exercise 39 below. However, we focus on the plus-equation and ask whether there are more solutions than those given above? With regard to (35) we have to ask for solutions of $Q_n = 1$. It can be shown that *Q_n is equal to one if and only if n is a multiple of the minimal period N*. The main idea behind is that the numbers $\alpha_0, \alpha_1, \ldots, \alpha_{N-1}$ are pairwise distinct (since N is the minimal period length), which implies that $Q_n = 1$ holds only whenever n is a multiple of N. However, for the rather technical proof we refer the interested reader to [38]. The solution $\mathcal{X}, \mathcal{Y} \in \mathbb{N}$ of (32) with minimal $\mathcal{X}$ is called **minimal solution**. By the remark from above, one can show that $(\mathcal{X}, \mathcal{Y}) = (x_1, y_1)$, where the right hand side is the solution of (32) according to (36). Anyway, it is clear that $\mathcal{X} \le p_{2N-1}$. Hence, solving the Pell equation is reduced to a finite problem. However, the length of periods of surds $\sqrt{d}$ can be rather long; see [31] for *fast* algorithms.

It is a well-known fact from algebra that the units of a ring form a multiplicative group. Therefore, it is not surprising, that the the solutions of the Pell equation define a group. In fact, we shall show that this group is cyclic, generated by the minimal solution. First, we return to our example with $d = 2$. We observe that

$$(\mathbf{3} + \mathbf{2}\sqrt{2})^2 = \mathbf{17} + \mathbf{12}\sqrt{2} \qquad \text{and} \qquad \mathbf{17}^2 - 2 \cdot \mathbf{12}^2 = (\mathbf{3}^2 - 2 \cdot \mathbf{2}^2)^2 = 1.$$

This leads to

Theorem 22 *The Pell equation has infinitely many solutions $x, y \in \mathbb{Z}$, all of them are up to the sign given by powers of the minimal solution $\mathcal{X}, \mathcal{Y} \in \mathbb{N}$:*

$$x + y\sqrt{d} := \pm(\mathcal{X} + \mathcal{Y}\sqrt{d})^{\pm n}, \quad \textit{where} \quad n = 0, 1, 2, \ldots.$$

Proof. In view of

$$\frac{\pm 1}{x+y\sqrt{d}} = \frac{\pm(x-y\sqrt{d})}{x^2-dy^2} = \pm x \mp y\sqrt{d} \tag{37}$$

it suffices to show that all solutions of (32) in positive integers are given by $x+y\sqrt{d} = \varepsilon^n$, where $\varepsilon := \mathcal{X}+\mathcal{Y}\sqrt{d}$ and $n \in \mathbb{N}$. For any positive solution we can find an $n \in \mathbb{N}$ such that

$$\varepsilon^n \le x+y\sqrt{d} < \varepsilon^{n+1}.$$

Define

$$X+Y\sqrt{d} = \varepsilon^{-n}(x+y\sqrt{d}),$$

then we have to prove that the latter expression is equal to one. Since $\sqrt{d}$ is irrational, conjugation (37) leads to

$$X-Y\sqrt{d} = \varepsilon^{n}(x-y\sqrt{d}).$$

Multiplying the latter equation with the previous one, we deduce

$$X^2-dY^2 = \varepsilon^{-n+n}\underbrace{(x-y\sqrt{d})(x+y\sqrt{d})}_{=x^2-dy^2=1} = 1.$$

Suppose now that $1 < X+Y\sqrt{d} < \varepsilon$, then, again with (37),

$$0 < \varepsilon^{-1} < (X+Y\sqrt{d})^{-1} = X-Y\sqrt{d} < 1.$$

It follows that

$$\begin{aligned} 2X &= (X+Y\sqrt{d})+(X-Y\sqrt{d}) > 1+\varepsilon^{-1} > 0, \\ 2Y\sqrt{d} &= (X+Y\sqrt{d})-(X-Y\sqrt{d}) > 1-1 = 0. \end{aligned}$$

Thus, X and Y are positive integral solutions of the Pell equation with $1 < X+Y\sqrt{d} < \varepsilon$. Since $x+y\sqrt{d}$ increases with y, we get $Y < \mathcal{Y}$ and $X < \mathcal{X}$, contradicting the fact that $\mathcal{X}, \mathcal{Y}$ is the fundamental solution. It follows that $X+Y\sqrt{d} = 1$. The assertion of the theorem follows. •

In the following table some continued fraction expansions of $\sqrt{d}$ and the related minimal solutions $(\mathcal{X}, \mathcal{Y})$ are listed:

$$\begin{array}{rcl} \sqrt{2} = [1,\overline{2}] & : & (3,2), \\ \sqrt{3} = [1,\overline{1,2}] & : & (2,1), \\ \sqrt{5} = [\overline{2,4}] & : & (9,4), \\ \sqrt{19} = [4,\overline{2,1,3,1,2,8}] & : & (170,39), \\ \sqrt{99} = [9,\overline{1,18}] & : & (10,1), \\ \sqrt{109} = [10,\overline{2,3,1,2,4,1,6,6,1,4,2,1,3,2,20}] & : & (158070671986249, 15140424455100), \\ \sqrt{2002} = [44,\overline{1,2,1,9,5,6,9,1,2,1,88}] & : & (11325887, 253128). \end{array}$$

It can be shown that the minimal solution of the Pell equation related to the cattle problem leads to an herd consisting of $N = 77602\ldots 81800$ many bulls and cows, where the number N has 206545 digits. On the webpage www.bioinfo.rpi.edu/ zukerm/cgi − bin/dq.html the interested (lazy?) reader can find an *online Pell solver.* The size of the minimal solution of the Pell equation behaves quite unregularly by increasing d. This is an interesting topic related to several deep and important questions; for example, the *class number problem.*

Another related topic are polynomial Pell equations. For instance, let $d = 1$ or $d = \pm 2$, then in [36] it is proved that *the sequences of polynomials given by* $P_0 = 1, Q_0 = 0$ *and, for* $n \in \mathbb{N}$*,*

$$\begin{aligned} P_n(X) &= \left(\frac{2}{d}X^2+1\right)P_{n-1}(X) + \frac{2}{d}X(X^2+d)Q_{n-1}(X), \\ Q_n(X) &= \frac{2}{d}XP_{n-1}(X) + \left(\frac{2}{d}X^2+1\right)Q_{n-1}(X) \end{aligned}$$

satisfy the polynomial equation $P(X)^2 - (X^2+d)Q(X)^2 = 1$, and that *these are (up to the sign) all solution of the latter equation.* However, it is an open problem to determine the polynomials D for which the polynomial Pell equation $P^2 - D\cdot Q^2 = 1$ has non-trivial solutions.

Since it has nothing to do with Pell or the Pell equation, we conclude with another interesting infinite representation due to Ramanujan:

Exercise 25 *Prove that*

$$n+1 = \sqrt{1+n\sqrt{1+(n+1)\sqrt{1+(n+2)\sqrt{\ldots}}}} \qquad \textit{for} \quad n \in \mathbb{N}.$$

10 Factorization with continued fractions

We are living in times with exponentially increasing electronic information exchange. This led to so-called *public key-cryptosystems* which mainly rely on so-called *one-way functions*, i.e. a simple operation which has an inverse that can hardly be computed. It is simple to multiply integers but, conversely, it is rather difficult to find the prime factorization of a given large integer; it is conjectured that the *factorization of integers is an* NP*-hard problem*, i.e., roughly speaking, there exists no *fast* algorithm. In the sequel we follow [10] and [28].

One of the first modern factorization methods is the **Continued fraction method** CFRAC due to Lehmer and Powers. This algorithm was first implemented by Brillhart and Morrison in 1970. As a proof for the power of it they computed the factorization of the 38-digit seventh **Fermat number** $F_7 := 2^{2^7}+1$; for the current knowledge on Fermat numbers see www.prothsearch.net/fermat.html♯Prime. In the following years CFRAC became the main factoring algorithm in practice; actually it was the first algorithm of *expected* sub-exponential running time. CFRAC relies in the main part on some ideas due to Fermat, Legendre and Kraitchik. Suppose that we are interested in the prime factorization of a large integer N. If there are integers x, y for which

$$x^2 \equiv y^2 \bmod N \qquad \text{and} \qquad x \not\equiv \pm y \bmod N,$$

then the greatest common divisor $(N, x+y)$ is a non-trivial factor of N. This follows immediately from the identity $x^2 - y^2 = (x-y)(x+y)$. To look randomly for pairs of squares which satisfy these conditions seems to be hopeless. The main idea is to collect sufficiently many squares which lie $\bmod N$ in the same residue class, such that certain combinations among them lead to non-trivial divisors of N. More precisely, having sufficiently many congruences

$$x_j^2 \equiv (-1)^{\varepsilon_{0j}} \ell_1^{\varepsilon_{1j}} \ell_2^{\varepsilon_{2j}} \cdot \ldots \cdot \ell_m^{\varepsilon_{mj}} \bmod N,$$

where the ℓ_k are *small* prime numbers and the ε_{kj} are the related exponents, by Gaussian elimination modulo 2 we may hope to find a relation of the form

$$\sum_{j \le n} \delta_j (\varepsilon_{0j}, \ldots, \varepsilon_{mj}) \equiv (0, \ldots, 0) \bmod 2, \tag{38}$$

where $\delta_j \in \{0, 1\}$. Then, setting

$$x = \prod_{j \le n} x_j^{\delta_j} \qquad \text{and} \qquad y = (-1)^{\nu_0} \ell_1^{\nu_1} \ell_2^{\nu_2} \cdot \ldots . \ell_m^{\nu_m}, \tag{39}$$

where $\sum_{j \le n} \delta_j (\varepsilon_{0j}, \ldots, \varepsilon_{mj}) = 2(\nu_0, \nu_1, \ldots, \nu_m)$, we get $x^2 \equiv y^2 \bmod N$. This splits N if $x \not\equiv \pm y \bmod N$. The set of prime numbers ℓ which are choosen to find the congruences in addition with -1 is called **factor basis**.

Whereas in the Quadratic sieve factoring algorithm the squares x_j^2 are generated by the nearest integers to $\sqrt{N}$ the continued fraction algorithm works with the numerators of the convergents to $\sqrt{N}$. This idea becomes more transparent in the algorithm

CFRAC

For $j = 0, 1, 2, \ldots$ successively:

1. Compute the jth convergent $\frac{p_j}{q_j}$ of the continued fraction expansion to $\sqrt{N}$.

2. Compute $p_j^2 \bmod N$ such that the absolute value is $\le \frac{1}{2} N$. After doing this for several j, look at the numbers $\pm p_j^2 \bmod N$ which factor into a product of small primes. Define your factor base $\mathcal{B}$ to consist of -1, the primes which either occur in more than one of the $p_j^2 \bmod N$ or which occur to an even power in just one $p_j^2 \bmod N$.

3. List all of the numbers $p_j^2 \bmod N$ which can be expressed as a product of numbers in the factor base $\mathcal{B}$. If possible, find a subset of numbers ℓ's of $\mathcal{B}$ for which the exponents ε according to the numbers in $\mathcal{B}$ sum to zero modulo two as in (38), and define x, y by (39). If $x \not\equiv \pm y \bmod N$, then $(x+y, N)$ is a non-trivial factor of N. If this is impossible, then compute more $p_j^2 \bmod N$, enlarging the factor base $\mathcal{B}$ if necessary.

Once the number of completely factored integers exceeds the size of the factor base, we can find a product of them which is a perfect square. With a little luck this yields a non-trivial factor of our given number (by the observations from above). The crucial property of the values p_j is, as we shall show below, that their squares have *small* residues modulo N. Otherwise, CFRAC would hinge on the problem of finding an appropriate factor base $\mathcal{B}$.

Theorem 23 *Let $\alpha > 1$ be irrational. Then the convertgents $\frac{p_n}{q_n}$ to α satisfy the inequality*

$$|q_n^2\alpha^2 - p_n^2| < 2\alpha.$$

In particular, if $\alpha = \sqrt{N}$, where $N \in \mathbb{N}$ is not a perfect square, then the residue $p_n^2 \bmod N$ which is smallest in absolute value is less than $2\sqrt{N}$.

Proof. In view of Theorem 11,

$$|q_n^2\alpha^2 - p_n^2| = q_n^2\left|\alpha - \frac{p_n}{q_n}\right| \cdot \left|\alpha + \frac{p_n}{q_n}\right| < \frac{q_n^2}{q_nq_{n+1}}\left(\alpha + \left(\alpha + \frac{1}{q_nq_{n+1}}\right)\right).$$

Thus,

$$|q_n^2\alpha^2 - p_n^2| - 2\alpha < 2\alpha\left(-1 + \frac{q_n}{q_{n+1}} + \frac{1}{2\alpha q_{n+1}^2}\right) < 2\alpha\left(-1 + \frac{q_n+1}{q_{n+1}}\right),$$

which is less or equal to zero, and proves the first assertion of the theorem. The claim on $p_n^2 \bmod N$ is an immediate consequence. •

Therefore, the sequence of the numerators of the convergents of $\sqrt{N}$ provides a sequence of p_j's whose squares have *small* residues $\bmod N$; in fact, for that one has even not to calculate the actual convergent in the continued fraction expansion, only the numerator p_j modulo N is needed.

We illustrate CFRAC by an example. Consider $N = 9073$. It is easy to compute the continued fraction expansion and its convergents:

$$\frac{95}{1}, \frac{286}{3}, \frac{381}{4}, \frac{10192}{107}, \frac{20765}{218}, \ldots \quad \rightarrow \quad \sqrt{9073} = [95, 3, 1, 26, 2, \ldots].$$

This leads to

j	0	1	2	3	4
$p_j \bmod 9073$	95	286	381	1119	2619
$p_j^2 \bmod 9073$	-48	139	-7	87	-27

It is not too difficult to factor the *small* numbers $p_j^2 \bmod 9073$. We may choose the factor base $\mathcal{B} = \{-1, 2, 3, 7\}$, so that $p_j^2 \bmod N$ is a $\mathcal{B}$-number for $j = 0, 2, 4$. The corresponding vectors of exponents α_j are

$$\{1, 4, 1, 0\}, \quad \{1, 0, 0, 1\} \quad \text{and} \quad \{1, 0, 3, 0\}.$$

The sum of the first and the third adds up to zero modulo two, which yields

$$x = 95 \cdot 2619 \equiv 3834 \bmod 9073 \qquad \text{and} \qquad y = 2^2 \cdot 3^2 = 36.$$

It follows that

$$3834^2 \equiv 36^2 \bmod 9073.$$

Since $3834 \not\equiv \pm 36 \bmod 9073$, we see that $(3834 + 36, 9073) = 43$ is a divisor of 9073 (the computation of greatest common divisors is easily done by the euclidean algorithms backwards). This leads to the prime factorization $9073 = 43 \cdot 211$.

Exercise 26 *Use* CFRAC *to factor the integers* $N = 17873, 25511$.

It is a well-known fact that CFRAC does not work for prime powers $N = p^k, k \geq 2$ (which is obvious if k is even but also true for odd exponents k). This is not a big problem. It is easy to check rather quickly whether a given N is a prime power or not. However there are other examples for which CFRAC does not work. If the period of the continued fraction expansion of $\sqrt{N}$ is too *short*, then the algorithm can produce only a *small* factor basis, which reduces the chances for factoring N. For instance, if $N = n^2 + 2$, then it is easily computed that $p_j^2 \bmod (n^2 + 2) = 1$ or $= -2$ (by hand or with our knowledge on the Pell equation with $d = N$). This gives in the CFRAC algorithm $(x + y, n^2 + 2) = n \pm 1$ which does not divide $n^2 + 2 = N$. A suitable refinement can be found in [54].

Having this powerful factoring algorithm in mind, we could be pessimistic about the security of cryptosystems:

Three may keep a secret, if two of them are dead. Benjamin Franklin

However, in modern cryptosystems prime numbers with approximately one hundred digits are used, whereas numbers which are the product of two such large primes cannot be factored in an appropriate time by the most modern factorization algorithms in practice. More details on this can be found in the excellent monography [10]; for applications of continued fraction algorithms to primality tests we refer to [7].

11 Liouville's theorem

A complex number α is called **transcendental** if it is not algebraic, i.e., there exists no polynomial with rational coefficients and root α. The use of the word *transcendental* dates at least back to Leibniz who wrote in 1704 *omnem rationem transcendunt.* First of all, it is not clear whether transcendental numbers exist. Using Cantor's notion of countability we can easily find an answer.

The **height** $H(P)$ of a polynomial P with integral coefficients is defined to be the maximum of all coefficients of P in absolute value. Let D, H be positive real numbers. Then it is easily seen that there exist only finitely many polynomials $P(X)$ with integral coefficients for which

$$\deg P \leq D \qquad \text{and} \qquad H(P) \leq H.$$

It follows that the set of polynomials with rational coefficients is countable. In particular, the set of roots of such polynomials is also a countable set. On the other side, $\mathbb{R}$ is uncountable. Thus

Theorem 24 *The set of algebraic numbers is countable, and the set of transcendental numbers is uncountable.*

This shows that almost all numbers are transcendental, but, actually, we do not know at least one transcendental number so far. This was one of the major problems from 19th century mathematics. In 1844 Liouville solved the problem.

Theorem 25 *For any algebraic number α of degree $d > 1$ there exists a positive constant c, depending only on α, such that*

$$\left|\alpha - \frac{p}{q}\right| > \frac{c}{q^d}$$

for all rationals $\frac{p}{q}$.

Proof. Denote by $P(X)$ the minimal polynomial of α. Then, by the mean-value theorem,

$$-P\left(\frac{p}{q}\right) = \underbrace{P(\alpha)}_{=0} - P\left(\frac{p}{q}\right) = \left(\alpha - \frac{p}{q}\right) P'(\xi)$$

for some ξ lying in between $\frac{p}{q}$ and α. Obviously, we may assume that $\left|\alpha - \frac{p}{q}\right| < 1$. Then $|\xi| < 1 + |\alpha|$ and hence $|P'(\xi)| < \frac{1}{c}$ for some positive c. It follows that

$$\left|\alpha - \frac{p}{q}\right| > c\left|P\left(\frac{p}{q}\right)\right|.$$

Since $P(X)$ is irreducible, it turns out that $\frac{p}{q}$ is not a zero of $P(X)$, and hence

$$\left|q^d P\left(\frac{p}{q}\right)\right| \geq 1,$$

which proves the theorem. •

The proof even provides a value for the constant c in terms of the height. The **height** $H(\alpha)$ of an algebraic number α is defined by $H(\alpha) = H(P_\alpha)$, where P_α is the minimal polynomial of α.

Exercise 27 *Prove that one can take*

$$c = \frac{1}{d^2(1+|\alpha|)^{d-1}H(\alpha)}$$

in Liouville's theorem.

Liouville's theorem shows that algebraic numbers cannot be approximated *too good* by rationals. Consequently, a real number which can be *better* approximated has to be transcendental! This is easily seen as follows: if we assume that $\alpha = [a_0, a_1, \ldots]$ is algebraic of degree d, then we have in view of equation (29) and Liouville's theorem the inequality

$$\frac{c}{q_n^d} < \left|\alpha - \frac{p_n}{q_n}\right| \leq \frac{1}{a_{n+1}q_n^2},$$

resp.

$$ca_{n+1} < q_n^{d-2}, \tag{40}$$

where the positive constant c depends only on α. If $d = 2$, the case of quadratic irrationals α, we see what we already know, namely that the sequence of partial denominators of α is bounded. But whenever

$$\liminf_{n\to\infty} \frac{a_{n+1}}{q_n} = \infty,$$

we may find for any positive δ infinitely many convergents to α for which $a_{n+1} \geq q_n^\delta$, contradicting (40). Hence, such an α cannot be algebraic. For instance, the real number $\alpha = [1, 10^{1!}, 10^{2!}, 10^{3!}, \ldots]$ is transcendental. However, it is more convenient to deal with series.

Exercise 28 *Show that*

$$\alpha = \sum_{n=1}^{\infty} 10^{-n!} = 0.11000\,10000\,\ldots$$

is transcendental.

In Exercise 5 we have proved that α is irrational.

A real number α is a **Liouville number** if for each positive integers m there exist numbers a_m, b_m such that

$$0 < \left|\alpha - \frac{a_m}{b_m}\right| < \frac{1}{b_m^m} \qquad \text{and} \qquad b_m > 1.$$

In view of Liouville's theorem it is easy to show, by the same reasoning as above, that *every Liouville number is transcendental.* It should be noted that Liouville wrote in his underlying paper

Je crois me souvenir qu'on trouve un théoréme de ce genre dans une lettre de Goldbach á Euler; mais je ne sache pas que la démonstration en ait jamais été donnée. (cf [8])

Liouville numbers are very interesting. One can prove that *Liouville numbers form an uncountable set of Lebesgue measure zero.* On the other side, Erdös [15] showed that *every real number can be written as a sum, resp. a product, of two Liouville numbers*; for this and other related results see [41].

12 The transcendence of e and π

The Liouville numbers from the previous paragraph look somehow artifical. It is much more difficult to decide whether a given number is transcendental or algebraic. The first and path-breaking result in this direction was proved by Hermite in 1873.

Theorem 26 *e is transcendental.*

Niven's proof for the irrationality of π (see Theorem 2) is based on Hermite's proof for the transcendence of e

Proof. Define

$$\mathcal{I}(t) = \int_0^t e^{t-x} f(x)\,\mathrm{d}x$$

for $t \geq 0$, where f is a real polynomial of degree m. Integration by parts shows that

$$\mathcal{I}(t) = -e^{t-x} f(x)\Big|_{x=0}^{t} + \int_0^t e^{t-x} f'(x)\,\mathrm{d}x = e^t \sum_{j=0}^{m} f^{(j)}(0) - \sum_{j=0}^{m} f^{(j)}(t). \tag{41}$$

Denote by F the polynomial obtained from f by replacing each coefficient with its absolute value. Then

$$|\mathcal{I}(t)| \leq \int_0^t e^{t-x} |f(x)|\,\mathrm{d}x \leq t e^t F(t). \tag{42}$$

Suppose now that e is algebraic, then there exists a polynomial P with integral coefficients a_k and leading coefficient $a_d \neq 0$ for which

$$\sum_{k=0}^{d} a_k e^k = 0.$$

Without loss of generality we may assume that P is the minimal polynomial of e. For a *large* prime p set

$$f(x) = x^{p-1} \prod_{k=1}^{d} (x-k)^p;$$

the degree of f is equal to $m = (d+1)p - 1$. We shall consider

$$\mathcal{J} := \sum_{k=0}^{d} a_k \mathcal{I}(k).$$

In view of (41)

$$\mathcal{J} = \sum_{k=0}^{d} a_k \left(e^k \sum_{j=0}^{m} f^{(j)}(0) - \sum_{j=0}^{m} f^{(j)}(k) \right) = -\sum_{j=0}^{m} \sum_{k=0}^{d} a_k f^{(j)}(k). \tag{43}$$

For $1 \le k \le d$ we have $f^{(j)}(k) = 0$ for $j < p$. Further, if $j \ge p$, then

$$f^{(j)}(k) = \binom{j}{p} p! \cdot g^{(j-p)}(k) \ , \qquad \text{where} \quad g(x) := \frac{f(x)}{(x-k)^p}.$$

Thus, for all j, $f^{(j)}(k)$ is an integer divisible by $p!$. Similarly, $f^{(j)}(0) = 0$ for $j < p-1$, and further, for $j \ge p-1$,

$$f^{(j)}(0) = \binom{j}{p-1} (p-1)! \cdot h^{(j-p+1)}(0) \ , \qquad \text{where} \quad h(x) := \frac{f(x)}{x^{p-1}}.$$

It follows that $h^{(j)}(0)$ is an integer divisible by p for $j > 0$, and $h(0) = (-1)^{dp}(d!)^p$. Consequently, for $j \ne p-1$, $f^{(j)}(0)$ is also an integer divisible by $p!$, and $f^{(p-1)}(0)$ is an integer divisible by $(p-1)!$ but not by p for $p > d$. Therefore, let $p > d$. It follows that J is a non-zero integer (by the first identity in (43)) which is divisible by $(p-1)!$ (by the second identity in (43)). Hence, $|\mathcal{J}| \ge (p-1)!$. On the other hand, the trivial bound $F(k) \le (2d)^m$ implies via the estimate (42)

$$|\mathcal{J}| \le \sum_{k=0}^{d} |a_k| k e^k F(k) \le H(e) d(d+1)(2d)^{(d+1)p-1} \le c^p,$$

where c is a constant independent of p. Thus,

$$(p-1)! \le |\mathcal{J}| \le c^p,$$

which is impossible for sufficiently large p. This contradiction shows that e cannot be algebraic. •

Following Hermite's ideas Lindemann succeeded in 1882 in showing

Theorem 27 *π is transcendental.*

Proof. Suppose that π is algebraic. It follows that also $\alpha := i\pi$ is algebraic, say of degree d. Denote the conjugates of α by $\alpha_1 = \alpha, \alpha_2, \ldots, \alpha_d$. From Euler's identity $\exp(i\pi) = -1$ it follows that

$$\underbrace{(1 + e^{\alpha_1})}_{=0}(1 + e^{\alpha_2}) \cdot \ldots \cdot (1 + e^{\alpha_d}) = 0.$$

The product on the left hand side can be written as a sum of 2^d terms e^{ϱ}, where

$$\varrho = \delta_1\alpha_1 + \ldots + \delta_d\alpha_d,$$

and $\delta_j = 0$ or 1. Now we assume that precisely n of the numbers ϱ are non-zero and denote them by $\beta_1, \ldots, \beta_n$. It follows that

$$q + e^{\beta_1} + \ldots + e^{\beta_n} = 0\ , \quad \text{where} \quad q := 2^d - n.$$

Now we shall proceed as in the previous proof and compare estimates for

$$\mathcal{H} := \sum_{k=1}^{n} \mathcal{I}(\beta_k),$$

where $\mathcal{I}(t)$ is defined by (41) with

$$f(x) = \ell^{np} x^{p-1} \prod_{k=1}^{n} (x - \beta_k)^p,$$

ℓ is the leading coeeficient of the minimal polynomial of α and p again denotes a large prime number, chosen later. We obtain

$$\mathcal{H} = -q \sum_{j=0}^{m} f^{(j)}(0) - \sum_{j=0}^{m} \sum_{k=1}^{n} f^{(j)}(\beta_k),$$

where $m = (n+1)p - 1$. Now the sum over k is a symmetric polynomial in $\ell\beta_1, \ldots, \ell\beta_n$ with integral coefficients. From the *fundamental theorem on symmetric functions* in addition with the observation that each elementary symmetric function in $\ell\beta_1, \ldots, \ell\beta_n$ is also an elementary symmetric function in the 2^d numbers $\ell\varrho$ it follows that the sum over k is an integer. Since $f^{(j)}(\beta_k) = 0$ for $j < p$ the sum in question is divisible by $p!$. Further, $f^{(j)}(0)$ is an integer divisible by $p!$ if $j \neq p-1$, and

$$f^{(p-1)}(0) = (p-1)!(-\ell)^{np}(\beta_1 \cdot \ldots \cdot \beta_n)^p,$$

which is an integer divisible by $(p-1)!$ but not by $p!$ when p is sufficiently large. Thus, $|\mathcal{H}| \geq (p-1)!$ if $p > q$. But on the other side, (42) implies the upper bound

$$|\mathcal{H}| \leq \sum_{k=1}^{n} |\beta_k| F(|\beta_k|) \leq c^p$$

for some constant c independent on p. As in the previous proof, this gives the contradiction. π is not algebraic. $\bullet$

As an important consequence of Lindemann's theorem one obtains a negative answer to an old and famous problem of the ancient Greek: *it is impossible to square a circle by using only ruler and compass.* This follows from the obvious fact that all points capable of construction are defined by intersection of lines and circles, and thus have only algebraic distances one from another; but on the other side, $\sqrt{\pi}$ is transcendental. Actually, Lindemann proved a stronger result, namely: *for any distinct algebraic numbers* $\alpha_1, \ldots, \alpha_n$ *and any non-vanishing algebraic numbers* $\beta_1, \ldots, \beta_n$,

$$\beta_1 e^{\alpha_1} + \ldots + \beta_n e^{\alpha_n} \neq 0.$$

The transcendence of e follows immediately, the one of π by Euler's identity.

Exercise 29 *Using Lindemann's theorem prove the transcendence of the trigonometric functions* $\sin\alpha, \cos\alpha$ *and* $\tan\alpha$ *for all non-zero algebraic* α.

There are many open problems in *transcendental number theory.* It is yet unproved whether e and π are *algebraically independent* or, to be more concrete, are $e+\pi$ and $e\cdot\pi$ both transcendental?

Exercise 30 *Show that at least one of the numbers* $e+\pi, e\cdot\pi$ *is transcendental.*

In his seventh problem Hilbert asked for a proof that α^β *is transcendental whenever* $\alpha \neq 0, 1$ *and* β *are algebraic irrationals.* In 1934 Gel'fond and Schneider (independently) succeeded in showing this deep result. A generalization was obtained by Alan Baker's celebrated estimates for linear forms in logarithms for which he was awarded with the Fields medal at the International Congress in Moscow 1966. For instance, he proved the transcendence of

$$\exp(\beta_0)\alpha_1^{\beta_1} \cdot \ldots \cdot \alpha_n^{\beta_n}$$

and that of any non-vanishing linear form

$$\beta_1 \log\alpha_1 + \ldots + \beta_n \log\alpha_n,$$

where the α_j and β_j denote non-zero algebraic numbers. For more details and a proof see [3], [45] or [8]

13 The theorem of Roth

It is natural to ask for stronger versions of Liouville's theorem. Only a slight improvement would imply the finiteness of integral solutions of certain important diophantine equations; a theme to which we shall return in a later section. Suppose that α is an algebraic number of degree d, then we might ask for exponents $\tau(d)$ such that there exists a positive constant c, depending only on α, such that for all rationals $\frac{p}{q}$

$$\left|\alpha - \frac{p}{q}\right| > \frac{c}{q^{\tau(d)}}. \tag{44}$$

The first refinement to Liouville's bound $\tau(d) = d$ was made by Thue who showed in 1909 that $\tau(d) > \frac{1}{2}d + 1$. In 1921 Siegel obtained $\tau(d) > 2\sqrt{d}$ (it should be noted that Siegel became professor in Frankfurt short after this discovery), and the famous physiscist Dyson proved in 1947 that $\tau(d) > \sqrt{2d}$. Finally, Roth proved in 1955 that $\tau(d)$ can be chosen independently of the algebraic degree d.

Theorem 28 *Let α be an algebraic number of degree $d \geq 2$ and let $\varepsilon > 0$. Then there exist only finitely many rational solutions $\frac{p}{q}$ to the inequality*

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

This shows that one can take any $\tau = \tau(d) > 2$ in (44); actually this is equivalent to Theorem 28. Consequently, algebraic numbers are approximable of order $\kappa = 2$ but not better. In view of Dirichlet's approximation theorem Roth's theorem is best possible with respect to the exponent 2. It might be possible that the ε-quantity can be sharpened. Lang conjectured that *for α of degree $d \geq 3$,*

$$\left|\alpha - \frac{p}{q}\right| > \frac{c}{q^2(\log q)^\kappa}$$

has only finitely many solutions if κ is a constant > 1. Roth got the Fields medal for his celebrated theorem during the International congress of mathematicians at Edinburgh in 1958. Unfortunately, Roth's theorem is ineffective. The method of proof does not provide an algorithm to determine the set of solutions. However, in particular cases it is possible to obtain upper bounds for the number of solutions (but this is a topic to which we will return later). A multidimensional generalization of Roth's theorem is the celebrated *subspace theorem* of W.M. Schmidt; a nice survey on this modern field of research is [43].

We can only sketch the idea of Roth's ingenious proof; the technical details are too complicated but the main ingredients become visible. We will follow [42] but first we review Liouville's idea. The proof of Theorem 25 depends mainly on

- the construction of a polynomial $P(X)$ with integral coefficients that vanishes at α and for which $P\left(\frac{p}{q}\right)$ is *small* in terms of $\left|\alpha - \frac{p}{q}\right|$, and
- the fact that $P\left(\frac{p}{q}\right)$ *cannot be too small* as a function of q.

Thue's ingenious contribution was to introduce polynomials in two variables which provides more freedom in the search for suitable polynomials. Before Roth it was already known that any progress would demand the use of polynomials in more than two variables. We start to formulate some properties which a polynomial should have. Suppose that for $1 \leq n \leq m$ the rational approximations $\frac{p_n}{q_n}$ to α satisfy

$$\left|\alpha - \frac{p_n}{q_n}\right| < \frac{1}{q_n^\tau}, \tag{45}$$

where τ is some positive real number. Let $P(X_1, \ldots, X_m)$ be a polynomial with integral coefficients, of degree at most r_n in X_n for each n. It follows that

$$\left|P\left(\frac{p_1}{q_1}, \ldots, \frac{p_m}{q_m}\right)\right| \geq \frac{1}{Q}\,, \qquad \text{where} \quad Q := q_1^{r_1} \cdot \ldots \cdot q_m^{r_m}, \tag{46}$$

provided that

$$P\left(\frac{p_1}{q_1},\ldots,\frac{p_m}{q_m}\right) \neq 0. \tag{47}$$

Let the Taylor expansion of $P\left(\frac{p_1}{q_1},\ldots,\frac{p_m}{q_m}\right)$ in powers of $\frac{p_n}{q_n}-\alpha$ be

$$\sum_{j_1}\cdots\sum_{j_m} c_{j_1,\ldots,j_m}\left(\frac{p_1}{q_1}-\alpha\right)^{j_1}\cdot\ldots\cdot\left(\frac{p_m}{q_m}-\alpha\right)^{j_m}.$$

Suppose now that P has in additon to (47) the properties that for small positive δ,

$$\sum_{j_1}\cdots\sum_{j_m}|c_{j_1,\ldots,j_m}| < Q^{\delta}, \tag{48}$$

and $c_{j_1,\ldots,j_m}=0$ for all $j_1,\ldots,j_m$ satisfying

$$q_1^{j_1}\cdot\ldots\cdot q_m^{j_m} \leq Q^{\Delta}, \tag{49}$$

where $\Delta>0$. Taking into account (45) we have for each term in the Taylor series with a non-zero coefficient

$$\left|\frac{p_1}{q_1}-\alpha\right|^{j_1}\cdot\ldots\cdot\left|\frac{p_m}{q_m}-\alpha\right|^{j_m} < \frac{1}{(q_1^{j_1}\cdot\ldots\cdot q_m^{j_m})^{\tau}} \leq \frac{1}{Q^{\tau\Delta}}.$$

Consequently,

$$\left|P\left(\frac{p_1}{q_1},\ldots,\frac{p_m}{q_m}\right)\right| < \frac{1}{Q^{\tau\Delta-\delta}}.$$

In view of (46) it follows that

$$\tau < \frac{1+\delta}{\Delta}. \tag{50}$$

In order to prove Theorem 28 we shall establish the existence of a polynomial P which satisfies the conditions (47), (48) and (49) with Δ near to $\frac{1}{2}$ as $\delta\to 0$. For this purpose it is necessary to take the number of variables m large and to choose suitable approximations $\frac{p_j}{q_j}$ (actually, taking $m=2$ leads only to an estimate $\tau < c\sqrt{d}$). In some sense, the question of good lower bounds for rational approximations to algebraic numbers is transformed to an approximation problem for polynomials in several variables; the use of many variables allows to put certain restrictions on the polynomial in question. We shall go a bit more into the details that at least the estimate $\tau>2$ becomes lucid.

The argument is indirect. We suppose that $\tau>2$. If there are infinitely many rationals satisfying (45), then we can find a subsequence $\frac{p_1}{q_1},\ldots,\frac{p_m}{q_m}$ for which

$$\frac{\log q_n}{\log q_{n-1}} > \frac{1}{\varepsilon}$$

with some positive ε and large q_1. Then we may choose positive integers r_n satisfying

$$q_1^{r_1} \leq q_n^{r_n} < q_1^{r_1(1+\frac{\varepsilon}{10})}.$$

Consequently, $q_1^{mr_1} \leq Q < q_1^{mr_1(1+\delta)}$. Then condition (49) takes the form that the Taylor coefficients $c_{j_1,\ldots,j_m}$ vanish for all $j_1,\ldots,j_m$ with

$$\frac{j_1}{r_1} + \ldots + \frac{j_m}{r_m} < m\Delta + \delta. \tag{51}$$

The next step is to construct a polynomial P^* that satisfies conditions (48) and (49) only. This is done following an argument of Siegel which bases on Dirichlet's pigeonhole principle. We put $B = q_1^{r_1}$ and consider the set of polynomials $W(X_1,\ldots,X_m)$ of degree at most r_n in X_n, and having integral coefficients, each of modulus less than B. We shall try to find two distinct polynomials W_1, W_2 such that their derivatives of order $j_1,\ldots,j_m$ are equal whenever $X_1 = \ldots = X_m = \alpha$, for all $j_1,\ldots,j_m$ satisfying (51). Any such derivative is of the form

$$a_0 + a_1\alpha + \ldots + a_{n-1}\alpha^n,$$

where the a_j are integers. It can be shown that the number of possibilities for a derivative for given $j_1,\ldots,j_m$ is $< B^{n(1+3\varepsilon)}$, whereas the number of polynomials W is about B^r, where $r = (r_1+1)\cdot\ldots\cdot(r_m+1)$. Hence, the number of possible distinct sets of derivatives provided that the number of sets $j_1,\ldots,j_m$ satisfying (51), and with no j_k exceeding the corresponding r_k, is less than about $\frac{r}{n(1+3\varepsilon)}$. The number of integer points $(j_1,\ldots,j_m)$ in the region defined by the above conditions can be shown to be less that $\frac{2r}{3n}$ if Δ is chosen by

$$m\Delta + \delta = \frac{1}{2}m - 3nm^{\frac{1}{2}}.$$

A more detailed analysis shows that we may assume that δ tends with ε to zero as $m \to \infty$. Thus we obtain

$$\Delta = \frac{1}{2} - \frac{\delta}{m} - \frac{3n}{m^{\frac{1}{2}}}.$$

This shows that the number of variables m has to be chosen large to give a value of τ nearby 2 in (50). This choice yields the existence of such polynomials W_1, W_2. Now the polynomial $P^* = W_1 - W_2$ satisfies condition (49), and it can be shown by some estimation process that it also satisfies (48). We cannot expect that P^* satisfies also condition (47). Actually, the construction of a polynomial P out of P^* which satisfies (47) is the most difficult aspect in Roth's theorem (Roth's lemma). Unfortunately, this is beyond the scope of this course; for details see the excellent [44], [45] and [60]. The final contradiction in Roth's proof is that for sufficiently large m we can find a Δ as close to $\frac{1}{2}$ as we please for which the right hand side of (50) exceeds any given $\tau > 2$.

There is an interesting correspondence to complex analysis due to Vojta. It is conjectured that the exponent 2 in Roth's theorem is the *same* number 2 which occurs in the *defect relation* in Nevanlinna's *value-distribution theory* or, roughly speaking, there seems to be a bridge between the approximation theories for numbers and for functions; for details see [59].

Roth's theorem has many important implications. For instance, following the lines of Liouville's proof of the transcendence of Liouville numbers, one can obtain a transcendence criterion for continued fractions.

Exercise 31 *Let $\varepsilon > 0$ and suppose that $\alpha = [a_0, a_1, \ldots]$ is an irrational number which has an infinity of partial quotients satisfying $a_{n+1} \geq q_n^{\varepsilon}$, where q_n is the denominator of the n-th convergent to α. Prove that α is transcendental.*

Another interesting consequence is Everett's attack on Fermat's last theorem. The main result of [16] states that *if p is a fixed odd prime and ν is a fixed positive integer, then there are at most finitely many pairs of coprime integers x, y on the line $y = x + \nu$ such that $x^p + y^p$ is the p-th power of an integer*, whereas in the case $p = 2$ *there are an infinitude of such pairs x, y for which $x^2 + y^2$ is a square*; the first result follows tricky but elementary from Theorem 28, and the second one from the theory of Pell equations. The most important implication of Roth's theorem is to find in the theory of cubic curves.

14 Thue equations

During his stay in London the ingenious Indian mathematician Ramanujan spend several weeks in hospital. Once his colleague Hardy came to visit and remarked that he had come in taxicab number 1729, which is *surely a rather dull number.* Ramanujan replied immediately that this is not true; 1729 is rather interesting since it is the smallest integer expressible as a sum of two cubes in two different ways:

$$1729 = 1^3 + 12^3 = 9^3 + 10^3.$$

This can be seen as follows. Since

$$(X + Y) \cdot (X^2 - XY + Y^2) = X^3 + Y^3 = 1729 = 7 \cdot 13 \cdot 19,$$

we have to consider all possible factorizations $1729 = A \cdot B$ and solve

$$A = X + Y \qquad \text{and} \qquad B = X^2 - XY + Y^2. \tag{52}$$

The substitution $Y = A - X$ leads to the quadratic equation

$$3X^2 - 3AX + A^2 - B = 0,$$

so that we only have to check if

$$\frac{1}{6}(3A \pm \sqrt{12B - 3A^2})$$

is an integer. This yields the two pairs $A = 13, B = 133$ and $A = 19, 91$ which correspond to the above given representations. We leave it to the reader to check that all integers $m < 1729$ have at most one representation as a sum of two cubes.

Exercise 32 *Find all integral solutions to the equation* $X^3 + Y^3 = 403$. *What can you say about the set of solutions of* $X^3 - Y^3 = m$ *with* $m = 0$ *and* $m = 1729$, *respectively?*

It is one thing to know that a certain diophantine equation has only finitely many solutions. However, for a complete list of all solutions one needs to have knowledge on their size.

Theorem 29 *Let* m *be a positive integer. Then every solution to the equation*

$$X^3 + Y^3 = m \tag{53}$$

in integers x, y *satisfies the inequality*

$$\max\{|x|, |y|\} \leq 2\sqrt{\frac{m}{3}}.$$

Proof. By the above reasoning each integral solution x, y satisfies (52) for some factorization $m = A \cdot B$. Hence

$$m \geq |B| = |x^2 - xy + y^2| = \frac{3}{4}x^2 + \left(\frac{1}{2}x - y\right)^2 \geq \frac{3}{4}x^2,$$

which gives the bound for x; since the equation is symmetric with respect to X and Y the same estimate holds for y as well. This proves the theorem. •

There is a big difference between the sets of rational numbers and of integers with respect to diophantine equations. We have shown that equation (53) has only finitely many *integral* solutions. On the contrary it can be shown that, for example, the equation

$$X^3 + Y^3 = 9$$

has infinitely many *rational* solutions. The transformation

$$X \mapsto \frac{12}{X + Y} \qquad \text{and} \qquad Y \mapsto 12\frac{X - Y}{X + Y}$$

provides a one-to-one correspondence between the rational solutions of the previous equation and the rational points on the curve

$$Y^2 = X^3 - 48;$$

the latter equation defines a so-called **elliptic curve** (this is not an ellipse but it is related to such objects). The *theory of elliptic curves* tells us that there are infinitely many rational solutions. The interested reader can find more details in the excellent monography [51]. A more simple but from a geometric point of view different example is the unit circle.

Analyzing (53) was rather simple since our diophantine equation was related to a reducible polynomial. If we have instead an irreducible polynomial, as for example $X^3 + 2Y^3 = m$, the situation is much more difficult. The first remarkable result for such equations was found by Thue in 1909.

Theorem 30 *Let a, b, c be non-zero integers. Then the equation*

$$aX^3 + bY^3 = c \tag{54}$$

has only finitely many solutions in integers.

Proof. If $x, y \in \mathbb{Z}$ is a solution to (54) then $X = ax, Y = y$ is a solution to $X^3 + a^2bY^3 = a^2c$. Thus it is enough to prove the theorem with $a = 1$. Further, since we may replace y by $-y$ and/or b by $-b$ in (54) if necessary, it is sufficient to consider the equation

$$X^3 - bY^3 = c, \tag{55}$$

where b, c are positive integers. The factorization method which we used in the proof of Theorem 29 worked very well. This observation motivates to have a look on the factorization

$$X^3 - bY^3 = (X - \beta Y) \cdot (X^2 + \beta XY + \beta^2Y^2) \qquad \text{with} \quad \beta = \sqrt[3]{3}. \tag{56}$$

If b is a perfect cube β, the latter identity is a factorization over the integers and we can factor c and proceed as above. If b is not a perfect cube, we need a different idea. Then we argue as Euler did in case of the Pell equation. If $x, y \in \mathbb{N}$ is a large solution of (55), then the first factor on the right hand side of (56) must have small modulus. This follows from the estimate

$$x^2 + \beta xy + \beta^2y^2 = \left(x + \frac{1}{2}\beta y\right)^2 + \frac{3}{4}(\beta y)^2 \geq \frac{3}{4}\beta^2y^2, \tag{57}$$

which implies via (56)

$$c = x^3 - by^3 = |x - \beta y| \cdot |x^2 + \beta xy + \beta^2y^2| \geq \frac{3}{4}\beta^2y^2 \cdot |x - \beta y|.$$

This gives

$$\left|\beta - \frac{x}{y}\right| \leq \frac{4c}{3\beta^2y^3}. \tag{58}$$

In view of Roth's theorem it follows that the latter inequality can have only finitely many solutions $\frac{x}{y}$ which implies the assertion of the theorem. •

Note that also Thue's estimate $\tau(3) > \frac{5}{2}$ in (44) is sufficient as well (Thue did not have Roth's bound) but not Liouville's $\tau(3) = 3$. Theorem 30 can easily be extended to equations of higher degree, so-called **Thue equations**.

Exercise 33 *Let $P(X, Y)$ be an irreducible binary form with integral coefficients of degree at least three and let m be any integer. Show that the equation $P(X, Y) = m$ has only finitely many solutions in integers. [Hint: instead of (57) work with a mean-value argument as in the proof of Liouville's theorem.]*

In view of Exercise 32 the condition on the irreducibility of P in the exercise above is necessary.

The approach via diophantine approximations has its limitations. Although the set of integral solutions is finite Thue's argument does not enable to give a complete list of solutions or indeed to determine whether the equation is soluble (we shall discuss this question in the following section). This follows from the fact that Roth's theorem is ineffective. In particular cases - here we do not mean *trivial* cases as settled in Theorem 29 - one can succeed by Baker's estimates for linear forms in logarithms. For example, Baker [2] obtained the estimate

$$\left|\frac{p}{q}-\sqrt[3]{2}\right|\geq\frac{10^{-6}}{q^{2.9955}},$$

valid for all rationals $\frac{p}{q}$. The exponent is larger than Roth's exponent $2+\varepsilon$ but the implicit constant is absolute. Combining this with (58) shows that any integer solution x,y to

$$X^3-2Y^3=c,$$

where c is a positive integer, satisfies the estimate

$$|y|\leq 10^{1317}c^{223}.$$

This is a large bound, but at least it only grows like a power of c.

Meanwhile more than Thue's theroem is known. The more general equation

$$aX^3+bX^2Y+cXY^2+dY^3+eX^2+fXY+gY^2+hX+iY+j=0$$

describes a **cubic curve**. In his PhD thesis Siegel proved that *a non-singular cubic curve with integral coefficients has only finitely many points with integral coordinates*; geometrically speaking, a point is called **singular** if it is a double point. It is easily seen that linear or quadratic equations can have an infinitude of solutions in integers (e.g. Pell's equation). Furthermore, by Faltings' celebrated proof of the Mordell conjecture we know that *any curve of genus* ≥ 2 *has only finitely many rational points*. Note that Faltings was awarded with a Fields medal at the International Congress of mathematicians in Berkeley 1986, following the footprints of Roth and Baker. We do not explain here what the topological notion *genus* means but note that this extends the class of homogeneous equations in two variables which Thue could handle and covers, for example, Fermat's last theorem (1). For a nicely written overview see the last chapter of [25].

15 The *abc*-conjecture

There is no *complete* theory for the variety of diophantine equations known. And it even seems that there exists nothing like that. In his tenth problem Hilbert asked if there exists a universal algorithm which can determine whether a given polynomial diophantine equation with integral coefficients has a solution in integers. In 1970 Matjasevic gave a negative answer; for the most simple proof see [26]. Besides he proved that *the set of prime numbers is diophantine*, where a set $\mathcal{A}$ is called **diophantine** if there exists a polynomial $P(X_1, \ldots, X_n)$ with integral coefficients such that the equation $P(X_1, \ldots, X_{n-1}, a) = 0$ has integral solutions if and only if $a \in \mathcal{A}$. On the other side, Putnam showed that a set is diophantine if and only if it coincides with the set of positive values of a suitable polynomial taken at the nonnegative integers. Consequently, *there exists a polynomial $g(X_1, \ldots, X_n)$ whose positive values at integers $x_j \geq 0$ are primes, and every prime can be represented this way.* Surprisingly, it is even possible to find such polynomials; here is an example due to Jones, Sato, Wada and Wiens of degree 25 in the 26 variables $a, b, c, \ldots, z$:

$$\begin{aligned}
&(k+2) \cdot \Big\{1 - [wz + h + j - q]^2 - [(gk + 2g + k + 1)(h + j) + h - z]^2 \\
&-[2n + p + q + z - e]^2 - [16(k+1)^3(k+2)(n-1)^2 + 1 - f^2]^2 \\
&-[e^3(e+2)(a+1)^2 + 1 - o^2]^2 - [(a^2-1)y^2 + 1 - x^2]^2 \\
&-[16r^2y^4(a^2-1) + 1 - u^2]^2 - [((a + u^2(u^2 - a))^2 - 1)(n + 4dy)^2 \\
&+1 - (x - cu)^2]^2 - [n + \ell + v - y]^2 \\
&-[(a^2-1)\ell^2 + 1 - m^2]^2 - [ai + k + 1 - \ell - i]^2 \\
&-[p + \ell(a - n - 1) + b(2an + 2a - n^2 - 2n - 2) - m]^2 \\
&-[q + y(a - p - 1) + s(2ap + 2a - p^2 - 2p - 2) - x]^2 \\
&-[z + p\ell(a - p) + t(2ap - p^2 - 1) - pm]^2\Big\}
\end{aligned}$$

(cf. [40]). It is not known which is the minimum possible number of variables is (clearly, it is larger than one), and it is also not known what the minimal degree is. With regard to Gödel's incompleteness theorem we finally note another remarkable consequence: *for any given axiomization there exists a diophantine equation without integral solutions (probably in three variables only) which cannot be proved in this axiomatic setting.*

Diophantine equations are a rather difficult topic. Of course, in special cases one might possibly succeed, sometimes even with *elementary* arguments.

Exercise 34 *Prove that $x = y = z = 0$ is the only integral solution of the equation*

$$X^3 + 2Y^3 + 4Z^3 - 34XYZ = 0.$$

[Hint: use **Fermat's method of descent**, *i.e. construct to a given (minimal) solution in positive integers a smaller one.]*

In particular cases it is much easier if we ask for a diophantine theory for polynomial equations (very often in number theory the function field cases are better to handle). Let $n(P)$ denote the number of distinct complex zeros of a polynomial P (which does not vanish identically). Recall that polynomials are said to be coprime if they do not share any of their linear factors, resp. if they have distinct roots. Stothers [58] and Mason [33] (independently) proved in the eighties of the last century the following surprising result.

Theorem 31 *Let a, b, c be coprime polynomials over $\mathbb{C}$, not all constant. If*

$$a + b = c,$$

then

$$\max\{\deg a, \deg b, \deg c\} < n(abc).$$

Proof. In view of the identity between the polynomials they are pairwise coprime, i.e., their zeros are pairwise distinct. By the *fundamental theorem of algebra* let

$$a = A\prod_{j=1}^{d_a}(X-\alpha_j)^{a_j}\ , \qquad b = B\prod_{k=1}^{d_b}(X-\beta_k)^{b_k} \quad \text{and} \quad c = C\prod_{l=1}^{d_c}(X-\gamma_l)^{c_l},$$

where A, B, C are complex numbers, d_a, d_b, d_c and a_j, b_k, c_l are nonnegative integers, and the $\alpha_j, \beta_k, \gamma_l$ are the distinct complex zeros of a, b and c, respectively. For later use we note the partial fraction decomposition of the logarithmic derivative

$$(\log a)' = \sum_{j=1}^{d_a}\frac{a_j}{X-\alpha_j},$$

where the derivative is taken with respect to X; obviously, analogous formulas hold for b and c, respectively. Without loss of generality we may assume that a has maximal degree; clearly, $\deg a \geq 1$. Setting $F = a/c$ and $G = b/c$ we obtain $F + G - 1 = 0$. This implies $F' + G' = 0$ and

$$\frac{a}{b} = \frac{F}{G} = -\frac{G'/G}{F'/F}.$$

We compute

$$\frac{F'}{F} = (\log F)' = (\log a)' - (\log c)' = \sum_{j=1}^{d_a}\frac{a_j}{X-\alpha_j} - \sum_{l=1}^{d_c}\frac{c_l}{X-\gamma_l},$$

and a similar formula holds for G'/G as well. This yields

$$\frac{a}{b} = -\frac{\displaystyle\sum_{k=1}^{d_b}\frac{b_k}{X-\beta_k} - \sum_{l=1}^{d_c}\frac{c_l}{X-\gamma_l}}{\displaystyle\sum_{j=1}^{d_a}\frac{a_j}{X-\alpha_j} - \sum_{l=1}^{d_c}\frac{c_l}{X-\gamma_l}}.$$

Multiplying both the denominator and the numerator with

$$\mathcal{K} := \prod_{j=1}^{d_a}(X-\alpha_j)\cdot\prod_{k=1}^{d_b}(X-\beta_k)\cdot\prod_{l=1}^{d_c}(X-\gamma_l),$$

leads to the representation

$$\frac{\mathcal{A}}{\mathcal{B}} = \frac{a \cdot \mathcal{K}}{b \cdot \mathcal{K}} = \frac{a}{b},$$

where $\mathcal{A}, \mathcal{B}$ are polynomials, both having degree $< d_a + d_b + d_c = n(abc)$ (termwise). Since a and b are coprime, and since polynomials over the complex numbers factor uniquely into irreducible (even in linear) factors, it follows that $\deg a < n(abc)$. The theorem is proved.

A slightly alternative proof relies on the *Riemann-Hurwitz formula* from *Algebraic geometry.* This idea of Silverman is outlined in [19] (which is also an excellent survey on the *abc*-conjecture). An immediate and interesting consequence of Theorem 31 is *Fermat's last theorem for polynomials.* Let $n > 2$ and suppose that we have polynomials x, y, z, not all constant, with

$$x^n + y^n = z^n.$$

Without loss of generality we may assume that x, y, z are coprime and that x has maximal degree. Put $a = x^n, b = y^n$ and $c = z^n$. Since the distinct zeros of $abc = (xyz)^n$ coincide with the distinct zeros of xyz, application of Theorem 31 yields

$$n \deg x = \deg a < n(abc) = n(xyz) \leq \deg xyz \leq 3 \deg x.$$

This proves that *all polynomial solutions of the Fermat equation (1) with exponent $n \geq 3$ are trivial, which here means that they are equal up to constant factors.* The condition on the exponent is necessary as the example

$$(2X)^2 + (X^2 - 1)^2 = (X^2 + 1)^2$$

shows; this should be compared with Exercise 1. It is remarkable that Fermat's last theorem for polynomials coincides with Fermat's last theorem for integers with respect to the exponents! *What can we learn out of this similarity?*

For another interesting application recall that, given a polynomial D, it is unknown how to determine whether the Pell equation $P^2 - DQ^2 = 1$ has polynomial solutions P, Q or not (see section 9). Taking into account Theorem 31 it can be shown that *if the number of distinct zeros of D is less than $\frac{1}{2} \deg D$, then the equation $P^2 - D \cdot Q^2 = 1$ is unsolvable.* For this and more see [56].

One approach to solve the lack of having a general approach towards a theory of diophantine equations is posing a general conjecture. First attempts were made by Oesterlé and in refined form by Masser soon after Theorem 31 appeared. Inspired by the above example of Fermat's last theorem one has to think about the correct translation of Theorem 31. By the *fundamental theorem of algebra* each polynomial over $\mathbb{C}$ splits into linear factors. So the degree of a polynomial is nothing else than the number of linear factors, and the distinct zeros correspond one-to-one to the distinct linear factors. The size of a number is measured by the absolute value and the irreducible factors of an integer are its prime factors. This dictionary is simple and sufficient. The most common version of the *abc*-**conjecture** states that *if a, b, c are coprime integers which satisfy*

$$a + b = c,$$

then, for any $\varepsilon > 0$,

$$\max\{|a|, |b|, |c|\} \leq C(\varepsilon) \left(\prod_{p|abc} p \right)^{1+\varepsilon},$$

where $C(\varepsilon)$ is a constant depending only on ε. The appearing product is called **squarefree kernel**. Unfortunately, the appearance of the ε and the related implicit constant are necessary as we shall explain now. Let m be a positive integer. *Fermat's little theorem* from elementary number theory states that *if a and n and are coprime integers, then*

$$a^{\varphi(n)} \equiv 1 \bmod n, \tag{59}$$

where $\varphi(n)$ is the order of the *group of prime residue classes* mod n $(\mathbb{Z}/p\mathbb{Z})^*$.This implies

$$2^{2\cdot 3^{m-1}} = 2^{\varphi(3^m)} \equiv 1 \bmod 3^m,$$

where . Hence, there exists an integer k for which

$$1 + 3^m k = 4^{3^{m-1}}.$$

Setting $a = 1, b = 3^m k$ and $c = 4^{3^{m-1}}$, it is immedately seen that

$$\mathcal{K} := \prod_{p|abc} p = 2 \cdot 3 \cdot \prod_{\substack{p|k \\ p\neq 2,3}} p \leq 6k.$$

It follows that

$$\frac{b}{\mathcal{K}} \geq 3^{m-2},$$

where the right hand side tends with m to infinity, whereas $\frac{b}{\mathcal{K}^{1+\varepsilon}}$ tends to zero if $\mathcal{K}$ is approximately of order k (which seems reasonable). In view of the rather general equation $a + b = c$ we have a plenty of interesting examples which all fit surprisingly well to the current state of knowledge on diophantine equations.

Theorem 32 *The abc-conjecture implies that there exist at most finitely many non-trivial solutions in integers of the Fermat equation*

$$X^n + Y^n = Z^n \qquad \textit{with} \quad n \geq 4.$$

Proof. If $x^n + y^n = z^n$ with coprime integers, then put $a = x^n, b = y^n$ and $c = z^n$ in the *abc*-conjecture. Without loss of generality we may assume that x, y and z

are positive integers. Since each prime factor of $abc = (xyz)^n$ divides xyz, their product is $\leq xyz < z^3$. Hence the abc-conjecture implies

$$z^n \leq C(\varepsilon) z^{3(1+\varepsilon)}.$$

In particular, if we choose $\varepsilon = \frac{1}{6}$ we get the estimate $z^{n-\frac{7}{2}} \leq C(\frac{1}{6})$ which leaves only finitely many possibilities for z if $n \geq 4$. This is the assertion. •

The proof suggests that the size of the exponents and the number of variables in diophantine equations (like the Fermat equation or also the Pell equation) seem to be crucial when we ask for solution in integers! Another example is the equation

$$X^m - Y^n = 1. \tag{60}$$

Recently, Mihailescu [34] proved the **Catalan conjecture** which claims that the only solution of the latter equation in integers $x, y, m, n \geq 2$ is given by $3^2 - 2^3 = 1$. It should be noted that Tijdeman proved via Baker's bounds for linear forms that there are only finitely many solutions.

Exercise 35 *Show that the abc-conjecture implies the finiteness of the set of solutions of (60) in integers whenever $m, n \geq 2$. What can be said about polynomial solutions?*

But the abc-conjecture can even do much more. Elkie [13] showed that *abc implies Mordell's conjecture* and Bombieri [6] deduced *Roth's theorem*; Frankenhuysen [17] unified both results and conjectured a refinement of Roth's statement (with respect to the appearance of ε in the eponent). Seeing how powerful the abc-conjecture is, we might be sceptic about its truth. On the contrary it produces exactly the results which are already proved or which are conjectured by a different reasoning. A proof of the abc-conjecture seems to be out of reach so far. There is some hope to generalize Wiles proof of Fermat's last theorem (which bases on an ingenious idea of Frey) along the theory of modular forms and Galois representations. Another attempt are again Baker's bounds for linear forms, but both approaches are not capable without new ideas. The best unconditional result (up to the appearing ε) is the estimate

$$\max\{|a|, |b|, |c|\} \leq \exp\left(C \prod_{p|abc} p^{\frac{1}{3}+\varepsilon}\right),$$

where the constant C is an effectively computable positive constant, due to Stewart and Kunrui [57] (using contributions by Waldschmidt and Tijdeman in advance).

16 p-adic numbers

p-adic numbers were discovered (or created, but this is a philosophical question) about hundred years ago by Hensel [22] (who was professor in Marburg). Meanwhile the theory of p-adic numbers has a plenty of applications and impacts in various mathematical fields but its origin lies in the theory of diophantine equations. We follow the monography [18] as well as Neukirch's survey in [12] and [55].

By the unique prime factorization of integers every rational number $\alpha \neq 0$ has a representation

$$\alpha = \pm \prod_p p^{\nu(\alpha;p)} \qquad \text{with} \quad \nu(\alpha;p) \in \mathbb{Z};$$

here the product is taken over all prime numbers p, but in fact only finitely many of the p-exponents $\nu(\alpha;p)$ of α are non-zero. Thus, if we fix a prime p, then we may write

$$\alpha = \frac{a}{b} \cdot p^{\nu(\alpha;p)} \qquad \text{with } 0 \neq a, b \in \mathbb{Z},\ p \not| ab.$$

Here and in the sequel p denotes always a prime number or the symbol ∞ which we will explain later. We define the **p-adic absolute value** on $\mathbb{Q}$ by setting

$$|\alpha|_p = \begin{cases} p^{-\nu(\alpha;p)} & \text{if} \quad \alpha \neq 0, \\ 0 & \text{if} \quad \alpha = 0; \end{cases}$$

the function $\alpha \mapsto \nu(\alpha;p)$ is called **p-adic valuation**. An **absolute value** on a field $\mathbb{K}$ is a function $|\ | : \mathbb{K} \to \mathbb{R}$ satisfying the axioms

- $|x| \geq 0$ for all $x \in \mathbb{K}$, and $|x| = 0$ if and only if $x = 0$;
- $|x \cdot y| = |x| \cdot |y|$ for all $x, y \in \mathbb{K}$;
- $|x + y| \leq |x| + |y|$ for all $x, y \in \mathbb{K}$.

If the last axiom can be replaced by

$$|x + y| \leq \max\{|x|, |y|\} \qquad \text{for all} \quad x, y \in \mathbb{K}, \tag{61}$$

the absolute value is said to be **non-archimedean**; otherwise the absolute value is called **archimedean**. The well-known absolute value

$$|\alpha|_\infty = \begin{cases} \alpha & \text{if} \quad \alpha \geq 0, \\ -\alpha & \text{if} \quad \alpha < 0, \end{cases}$$

is the standard example of an archimedean absolute value on $\mathbb{K}$; the notation $|\ |_\infty$ is traditional in the context of p-adic numbers. An example of a non-archimedean absolute value is the **trivial** absolute value which is constant 1 on all non-zero elements, but this is boring. More interesting are p-adic absolute values. It is easy to check that $|\ |_p$ is indeed an absolute value, and it is not much more difficult to prove that even (61) is satisfied; the main idea for the proof can be found by studying the following example

$$|3 \cdot 5 + 2 \cdot 3^2|_3 = |3(5 + 2 \cdot 3)|_3 = \frac{1}{3} = \max\{|3 \cdot 5|_3, |2 \cdot 3^2|_3\}.$$

Exercise 36 *Prove that the p-adic absolute is a non-archimedean absolute value. Show that an absolute value $|\ |$ on a field $\mathbb{K}$ is non-archimedean if and only if*

$$\sup\{|n| \,:\, n \in \mathbb{N}\} < \infty; \tag{62}$$

Remark: note that, since any positive integer n has a representation $n = 1 + \ldots + 1$, $\mathbb{N}$ has a natural embedding into any field $\mathbb{K}$.

Taking into account characterization (62) we see that the origin of the notion of an archimedean absolute value is caused by *Archimedes' lemma* which states that *for all non-zero integers, resp. rationals x, y there exists a positive integer m with*

$$|mx|_\infty > |y|_\infty.$$

An absolute value on a field $\mathbb{K}$ induces a topology. Since an absolute value $|\ |$ is a norm, we may define a metric by setting

$$\mathrm{d}(x, y) = |x - y| \qquad \text{for } x, y \in \mathbb{K}.$$

Now we can measure distances on $\mathbb{K}$. If $|\ |$ is a non-archimedean absolute value, then we call the corresponding metric **ultrametric** and $\mathbb{K}$ together with this ultrametric is said to be an **ultrametric space**. Obviously, with view to (61), a metric is ultrametric if and only if

$$\mathrm{d}(x, y) \leq \max\{\,\mathrm{d}(x, z),\, \mathrm{d}(y, z)\} \qquad \text{for all } x, y, z \in \mathbb{K}$$

holds; the latter inequality is called the **ultrametric inequality**. We note that a rational number α has a small p-adic absolute value if and only if α is divisible by a large power of p. This was the underlying idea for Hensel in introducing p-adic numbers. Divisibility properties of the integers are the fundamentals in number theory!

We call two absolute values **equivalent** if they induce the same topology. It is natural to ask what kind of absolute values a given field has. In 1918 Ostrowski gave a full description of the absolute values of the field of rational numbers; he proved that *every non-trivial absolute value on $\mathbb{Q}$ is equivalent to one of the absolute values $|\ |_p$, where either p is a prime number or $p = \infty$*. This fits perfectly to the **product formula**

$$\prod_{p \leq \infty} |\alpha|_p = 1 \qquad \text{for any} \quad 0 \neq \alpha \in \mathbb{Q};$$

the latter identity follows easily from the definition of the p-adic absolute value and the unique prime factorization. The primes p are also called **places** and $p = \infty$ stands for the **infinite place**.

p-adic absolute values imply a curious convergence. Consider the linear equation

$$X = pX + 1.$$

It is a simple task to find the solution by separating all X-terms. But we shall try something completely different. The iteration

$$x_0 := 1\ , \quad x_n := px_{n-1} + 1 \qquad \text{for} \quad n = 1, 2, 3, \ldots$$

leads to the sequence

$$x_n = 1 + p + p^2 + \ldots + p^n = \frac{1 - p^{n+1}}{1 - p}.$$

Since

$$\left| \frac{p^{n+1}}{1-p} \right|_p = |p^{n+1}|_p \cdot |1 - p|_p^{-1} = p^{-n-1}$$

tends to zero as $n \to \infty$, we see that the sequence (x_n) is p-adically convergent. Moreover, we obtain a strange formula for the geometric series

$$\sum_{k=0}^{\infty} p^k = \frac{1}{1-p}, \tag{63}$$

which is with respect to the usual absolute value divergent! Actually, this is the solution of the equation in question (since the solution, resp. the limit, is rational we are out of any trouble). However, the same reasoning as for (63) can be applied to other series as well:

$$\alpha = \sum_{k \geq \nu} a_k p^k \qquad \text{with} \quad \alpha_k \in \mathbb{Z},\ 0 \leq a_k < p, \tag{64}$$

where $\nu \in \mathbb{Z}$ (but $\nu = -\infty$ is forbidden), are p-adically convergent. Moreover, the sequences of their partial sums are all Cauchy sequences. This follows immediately from the following assertion. *A sequence of rational numbers (α_n) is a Cauchy sequence with respect to a p-adic absolute value $|\ |_p$ if and only if*

$$\lim_{n \to \infty} |\alpha_{n+1} - \alpha_n|_p = 0. \tag{65}$$

This gives a first glimpse on p-adic analysis and shows how much it differs from real analysis. To prove (65) we write $m = n + r > n$, and note with regard to (61) that

$$\begin{aligned} |\alpha_m - \alpha_n|_p &= |\alpha_{n+r} \underbrace{-\alpha_{n+r-1} + \alpha_{n+r-1}}_{=0} - \alpha_{n+r-2} \pm \ldots + \alpha_{n+1} - \alpha_n|_p \\ &\leq \max\{|\alpha_{n+r} - \alpha_{n+r-1}|_p, \ldots, |\alpha_{n+1} - \alpha_{n+r-1}|_p\}. \end{aligned}$$

Similarly to the decimal fraction expansion we have the following irrationality criterion:

Exercise 37 *Show that the series (64) has a rational limit if and only if the sequence of the* **ciphers** *a_k is eventually periodic.*

We know that $\mathbb{Q}$ is not complete with respect to the usual archimedean absolute value (e.g., Theorem 1). Actually, by the latter exercise the same is true if we take any p-adic absolute value. Thus the field of rational numbers is not complete to any of its non-trivial absolute values. To get out of this trouble we may complete $\mathbb{Q}$ with respect to a p-adic absolute value, analogously to Cantor's construction of the real numbers as a completion of $\mathbb{Q}$ with respect to $|\ |_p$. Denote by $\mathcal{C}_p$ the set of all p-adic Cauchy sequences (α_n). We define addition and multiplication by

$$(\alpha_n)+(\beta_n)=(\alpha_n+\beta_n) \qquad \text{and} \qquad (\alpha_n)\cdot(\beta_n)=(\alpha_n\cdot\beta_n).$$

This makes $\mathcal{C}_p$ to a commutative ring. The subset $\mathcal{M}_p$ consisting of all Cauchy sequences (α_n) with $\lim_{n\to\infty}\alpha_n=0$ is an ideal (that's clear). In fact, $\mathcal{M}_p$ is even a maximal ideal which can be seen as follows.

Suppose we have an arbitrary Cauchy sequence $(\alpha_n)\in\mathcal{C}_p\setminus\mathcal{M}_p$, i.e., $\lim_{n\to\infty}\alpha_n\neq 0$. Then there exist a constant c and an integer N such that

$$|\alpha_n|_p\geq c>0 \qquad \text{for} \quad n\geq N.$$

Define (β_n) by setting $\beta_n=0$ if $n<N$, and $\beta_n=\frac{1}{\alpha_n}$ otherwise. For $n\geq N$,

$$|\beta_{n+1}-\beta_n|_p=\left|\frac{1}{\alpha_{n+1}}-\frac{1}{\alpha_n}\right|_p=\left|\frac{\alpha_{n+1}-\alpha_n}{\alpha_{n+1}\alpha_n}\right|_p\leq\frac{1}{c^2}|\alpha_{n+1}-\alpha_n|_p.$$

Hence, it follows from (65) that (β_n) is a Cauchy sequence since (α_n) is one. Note that $\lim_{n\to\infty}\alpha_n\beta_n=1$, and thus

$$(1)-(\alpha_n)\cdot(\beta_n)\in\mathcal{M}_p.$$

This shows that the ideal which is generated by (α_n) and $\mathcal{M}_p$ is equal to the ideal generated by (1), but that is the whole ring $\mathcal{C}_p$. This shows that $\mathcal{M}_p$ is maximal.

From algebra we know that the quotient of a commutative ring by its maximal ideal is a field.

Theorem 33 $\mathbb{Q}_p:=\mathcal{C}_p/\mathcal{M}_p$ *is a field, the* **field of p-adic numbers**.

We can embed $\mathbb{Q}$ via the mapping $\alpha\mapsto(\alpha,\alpha,\ldots)$ in a natural way into $\mathbb{Q}_p$, and thus $\mathbb{Q}_p$ is the completion of $\mathbb{Q}$ with respect to $|\ |_p$. Obviously, $\mathbb{Q}$ lies dense in $\mathbb{Q}_p$. The p-adic absolute value can be continued from $\mathbb{Q}$ onto $\mathbb{Q}_p$. In view of Ostrowski's theorem, the non-equivalent non-trivial absolute values on $\mathbb{Q}$ lead to a family of fields lying above $\mathbb{Q}$:

$$\mathbb{Q} \quad \overset{\text{completion}}{\longrightarrow} \quad \mathbb{Q}_2,\mathbb{Q}_3,\mathbb{Q}_5,\cdots \quad \text{and} \quad \mathbb{Q}_\infty=\mathbb{R}.$$

Nearby the world of real numbers exist - with the same right - for each prime number p the world of p-adic numbers.

But how do p-adic numbers look like? Actually, we know p-adic numbers already. Recall the series representations (64). By (65) these series are the limits of Cauchy sequences of rational numbers. We may identify them with elements in $\mathbb{Q}_p$, i.e., each series (64) defines a p-adic number α, and any p-adic number α has a representation of the form (64). The first assertion is clear, the second one can be shown by an approximation argument as follows. Firts, suppose that α is a p-adic number with $\nu(\alpha;p) \geq 0$ and n is a positive integer. Since $\mathbb{Q}$ is dense in $\mathbb{Q}_p$ we can find a rational number $\frac{x}{y}$ such that

$$\left|\alpha - \frac{x}{y}\right|_p \leq p^{-n}.$$

In view of (61)

$$\left|\frac{x}{y}\right|_p = \left|\frac{x}{y}\underbrace{-\alpha+\alpha}_{=0}\right|_p \leq \max\left\{|\alpha|_p, \left|\alpha - \frac{x}{y}\right|_p\right\} \leq 1.$$

Consequently, p does not divide y. Hence there exists an integer Y_n such that $yY_n \equiv 1 \bmod p^n$. Moreover, by (8) we may assume that $0 \leq xY_n \leq p^n - 1$. This implies

$$\left|\frac{x}{y} - xY_n\right|_p = \left|\frac{x - xyY_n}{y}\right|_p \leq p^{-n}.$$

Put $\alpha_n = xY_n$. Then, with regard to the above estimates,

$$|\alpha - \alpha_n|_p = \left|\alpha \underbrace{-\frac{x}{y} + \frac{x}{y}}_{=0} - \alpha_n\right|_p \leq \max\left\{\left|\alpha - \frac{x}{y}\right|_p, \left|\frac{x}{y} - \alpha_n\right|_p\right\} \leq p^{-n}.$$

By induction it follows that for any such α there exists a Cauchy sequence of integers (α_n) converging to α for which

$$0 \leq \alpha_n \leq p^n - 1 \qquad \text{and} \qquad \alpha_n \equiv \alpha_{n-1} \bmod p^{n-1}.$$

This implies the existence of a sequence of integers a_k with

$$\alpha_n = a_0 + a_1 p + \ldots + a_n p^n \qquad \text{and} \quad 0 \leq a_k \leq p-1, \tag{66}$$

which leads to a representation of α of the form (64); the general case is deduced form the one above by considering $\alpha p^{-\nu(\alpha;p)}$ instead of α.

Hensel introduced p-adic numbers in an ad hoc manner via (64). His idea was to transport the powerful method of power series from analysis to number theory. This direct approach is of advantage for explicit computations.

Exercise 38 *Calculate the 3-adic expansions of $\frac{5}{9}, \frac{9}{5}, \frac{5}{9} + \frac{9}{5}$ and $\frac{5}{9} \cdot \frac{9}{5}$. What is the p-adic value of*

$$\sum_{k=0}^{\infty} (p-1)p^k \ ?$$

We shall give a second, purely algebraic construction of $\mathbb{Q}_p$ which we will use later on. Denote by $\mathbb{Z}/p^n\mathbb{Z}$ the ring of residue classes $\bmod p^n$ for $n \in \mathbb{N}$. Then we may define the sequence of maps

$$\ldots \to \mathbb{Z}/p^{n+1}\mathbb{Z} \to \mathbb{Z}/p^n\mathbb{Z} \to \ldots \to \mathbb{Z}/p^2\mathbb{Z} \to \mathbb{Z}/pZ,$$

where each map is the natural projection

$$\pi_n \,:\, \mathbb{Z}/p^{n+1}\mathbb{Z} \to \mathbb{Z}/p^n Z \;, \quad x \mapsto x \bmod p^n.$$

Now we can define the **projective limit**

$$\lim_{\leftarrow} \mathbb{Z}/p^n\mathbb{Z} = \left\{ (x_n) \in \prod_{n \geq 1} \mathbb{Z}/p^n\mathbb{Z} \,:\, \pi(x_{n+1}) = x_n \right\}.$$

Being a formal product of rings the projective limit inherits the ring structure of its factors. As a matter of fact we see that each $(x_n) \in \lim_{\leftarrow} \mathbb{Z}/p^n\mathbb{Z}$ gives raise to a Cauchy sequence of integers α_n (with respect to the p-adic absolute value) such that

$$x_n \equiv \alpha_n \bmod p^n,$$

resp. a sequence of integers (a_k) satisfying (66). It follows that

$$\lim_{\leftarrow} \mathbb{Z}/p^n\mathbb{Z} \cong \left\{ \sum_{k=0}^{\infty} a_k p^k \,:\, 0 \leq a_k \leq p-1 \right\}. \tag{67}$$

Denote the right hand side above by $\mathbb{Z}_p$. With the projective limit also $\mathbb{Z}_p$ is a commutative ring. Now we can define $\mathbb{Q}_p$ as the fraction field of $\mathbb{Z}_p$. According to this construction we say that a p-adic number α with $|\alpha|_p \leq 1$, resp. $\nu(\alpha; p) \geq 0$, is a **p-adic integer**, and $\mathbb{Z}_p$ becomes the **ring of p-adic integers**.

17 The Local-global principle

The p-adic way of thinking gives a new strategy to attack diophantine equations. Consider the equation

$$P(X_1, X_2, \ldots, X_r) = 0, \tag{68}$$

where P is a polynomial in several variables with integral coefficients. We are interested in the solubility in integers. We can weaken this difficult problem by replacing (68) by the system of congruences

$$P(X_1, X_2, \ldots, X_r) \equiv 0 \bmod m,$$

where m runs through all positive integers, or equivalently, by the *chinese remainder theorem*,

$$P(X_1, X_2, \ldots, X_r) \equiv 0 \bmod p^n \qquad \text{with} \quad n = 1, 2, \ldots, \tag{69}$$

where p runs through all prime numbers. This approach yields sometimes certain information about the original equation. We shall illustrate this by an interesting example. We ask for the integral solutions of

$$Y^2 = X^3 + 7.$$

By Exercise 33 we know that there are at most finitely many solutions in integers. We may rewrite the equation in question as

$$Y^2 + 1 = (X + 2)((X - 1)^2 + 3).$$

If $x, y \in \mathbb{Z}$ is a solution, then $(x-1)^2 + 3 \equiv 3 \bmod 4$. Hence there is a prime $p \equiv 3 \bmod 4$ dividing it, and reducing the equation in question modulo p shows that -1 is a square $\bmod p$ which gives a contradiction. Thus, there are no integer solutions to the equation in question.

Exercise 39 *Show that the Pellian minus-equation*

$$Y^2 - dY^2 = -1$$

has no integral solutions if $d \equiv 3 \bmod 4$.

Up till now we made out of one equation infinitely many congruences. But we can go further. It is a very astonishing and useful fact that all these unpleasantly many congruences (69) for a fixed prime can be summed up to one p-adic equation.

Theorem 34 *With the conditions on P from above, the system of congruences (69) is solvable if and only if the equation (68) has a solution in p-adic integers.*

Proof. Via the projective limit (67) we have

$$\mathbb{Z}_p \cong \lim_{\leftarrow} \mathbb{Z}/p^n\mathbb{Z} \subset \prod_{n=1}^{\infty} \mathbb{Z}/p^n\mathbb{Z}.$$

The equation (68) splits over the ring on the right hand side into the system of congruences (69). Thus, each p-adic solution of (68) leads also to a solution of (69).

Conversely, for any positive integer n, let $(x_1^{(n)}, x_2^{(n)}, \ldots, x_r^{(n)})$ be a solution of

$$P(X_1, X_2, \ldots, X_r) \equiv 0 \bmod p^n.$$

If all elements

$$(x_1^{(n)})_n, (x_2^{(n)})_n, \ldots, (x_r^{(n)})_n \in \prod_{n=1}^{\infty} \mathbb{Z}/p^n\mathbb{Z}$$

lie in $\lim_{\leftarrow} \mathbb{Z}/p^n\mathbb{Z} = \mathbb{Z}_p$, we are ready. Otherwise, we have to construct a subsequence with this property. We only consider the case $r = 1$ and write x instead of x_1; the general case can be treated similarly. Since $\mathbb{Z}/p\mathbb{Z}$ is finite there are infinitely many terms of $x^{(n)}$ which are congruent to some fixed $y_1 \in \mathbb{Z}/p\mathbb{Z}$. Thus we can find a subsequence $\{x_1^{(n)}\}$ of $\{x^{(n)}\}$ for which

$$x_1^{(n)} \equiv y_1 \qquad \text{and} \qquad P\left(x_1^{(n)}\right) \equiv 0 \bmod p.$$

Obviously, we can continue and obtain, for any $k \geq 2$, a subsequence $\{x_k^{(n)}\}$ of $\{x_{k-1}^{(n)}\}$ such that

$$x_k^{(n)} \equiv y_k \qquad \text{and} \qquad P\left(x_k^{(n)}\right) \equiv 0 \bmod p^k,$$

where the $y_k \in \mathbb{Z}/p^k\mathbb{Z}$ are related by

$$y_k \equiv y_{k-1} \bmod p^{k-1}.$$

The y_k define a p-adic integer $y = (y_k)_k \in \lim_{\leftarrow} \mathbb{Z}/p^k\mathbb{Z} \cong \mathbb{Z}_p$ satisfying

$$P(y_k) \equiv 0 \bmod p^k \qquad \text{for every} \quad k \in \mathbb{N}.$$

By continuity this implies $P(y) = 0$. The theorem is proved. •

If the polynomial P is homogeneous, the equation (68) has always the trivial solution $x_1 = \ldots = x_r = 0$. In this context it is more natural to ask for non-trivial solutions, i.e., solutions where not all x_j equal zero. In this case a little variation of the proof above gives the corresponding statement for a non-trivial p-adic solution.

We shall give an example for Theorem 34. Consider the congruences

$$X^2 \equiv 2 \bmod 7^n \qquad (n = 1, 2, \ldots). \tag{70}$$

Since 2 is a quadratic residue mod 7 the congruence is solvable for $n = 1$. Indeed, we find thesolutions $\pm 3 \bmod 7$. We start with $x_1 \equiv +3 \bmod 7$. Now let $n = 2$. Any solution x_2 of (70) with $n = 2$ satisfies also (70) with $n = 1$. We do the ansatz $x_2 \equiv 3 + 7z$ which leads in (70) to

$$2 \equiv (3 + 7z)^2 = 9 + 6z \cdot 7 + 7^2 z^2 \equiv 2 + (1 + 6z) \cdot 7 \bmod 7^2,$$

resp.

$$0 \equiv 1 + 6z \bmod 7.$$

Thus, $z \equiv 1 \bmod 7$ and we obtain $x_2 \equiv 3 + 1 \cdot 7 \bmod 7^2$. Since in any further step we have to solve only linear congruences mod 7, this process continues ad infinitum and leads to

$$x = 3 + 1 \cdot 7 + 2 \cdot 7^2 + \ldots,$$

which is a 7-adic solution of the equation $X^2 = 2$; we may write $x = \sqrt{2}$ but notice that this is not the square root of 2 in the field of real numbers. The other solution can be found the same way starting with $x_1 \equiv -3 \bmod 7$. There is another aspect in this example which is of certain interest. The crucial step above was the fact that 2 is a quadratic residue mod 7, linear congruences are always solvable. This observation shows that *a positive integer α is a square in $\mathbb{Z}_p$ if and only if α is a quadratic residue* mod p. Furthermore, we have

Theorem 35 *A rational number α is a square if and only if it is a square in all $\mathbb{Q}_p$ for all $p \leq \infty$.*

Proof. One implication follows easily from the embedding $\mathbb{Q} \subset \mathbb{Q}_p$. For the other one we have a look on the prime factorization of α:

$$\alpha = \pm \prod_{p<\infty} p^{\nu(\alpha;p)}.$$

If α is a square in the field of real numbers, α is positive. Furthermore, if α is a square in $\mathbb{Q}_p$, the exponent $\nu(\alpha;p)$ is even. This proves the other implication. •

Theorem 35 has interesting consequences for the structure of p-adic fields.

Exercise 40 *Show that two distinct p-adic fields are non-isomorphic. Further, prove that $\mathbb{R}$ and any $\mathbb{Q}_p$ are non-isomorphic.*

Finite extensions of $\mathbb{Q}$ are called **global**, completions of global fields with discrete valuation and finite residue field are **local**. The so-called **local-global principle** is the idea of *putting together information from all local fields $\mathbb{Q}_p$ and additionally $\mathbb{R} = \mathbb{Q}_\infty$, to get information in the global field $\mathbb{Q}$*. This principle is extraordinary successful and seems to go back to Hensel but was first clearly stated by Hasse. Theorem 35 is only the very beginning but gives a first glimpse of its power. With this concept Hasse was able to give around 1920 an important characterization of so-called isotropic quadratic forms. Let $\mathbb{K}$ be a field. Then we say that a quadratic form $P \in \mathbb{K}[X_1, \ldots, X_r]$ is **isotropic** over $\mathbb{K}$ if there exist $x_1, \ldots, x_r \in \mathbb{K}$, not all $x_k = 0$, such that

$$P(x_1, \ldots, x_r) = 0.$$

It can be shown that *isotropic quadratic forms take all values, i.e., for any $\beta \in \mathbb{K}$, the equation*

$$P(X_1, \ldots, X_r) = \beta$$

has a solution in $\mathbb{K}$. A first but unsatisfying classification of isotropic quadratic forms was found by Minkowski in 1890. He proved that *if $P(X_1, \ldots, X_r)$ is a quadratic form with integer coefficients for which (69) with any prime p has a non-trivial solution, and if $P(X_1, \ldots, X_r)$ has a non-trivial real solution, then P is isotropic over $\mathbb{Q}$*. In view of Theorem 34 this can be translated into the celebrated theorem of Hasse-Minkowski which gives a particular solution of Hilbert's eleventh problem on representations of integers in algebraic number fields by quadratic forms.

Theorem 36 *A quadratic form is isotropic over $\mathbb{Q}$ if and only if it is isotropic over all $\mathbb{Q}_p, p \leq \infty$.*

Hasse's p-adic proof is much easier and more natural than Minkowski's approach, but anyway even Hasse's proof is far beyond the scope of these notes (it uses deeper knowledge on the *theory of quadratic forms*, the multiplicative structure of $\mathbb{Q}_p$ and its subgroup of squares, and even *Dirichlet's prime number theorem for arithmetic progressions*). We refer the interested reader to [50].

It is easily seen that the ring of p-adic units in $\mathbb{Z}_p$ is given by

$$\mathbb{Z}_p^* := \{\alpha \in \mathbb{Q}_p \,:\, |\alpha|_p = 1\}.$$

One can show that *if p is an odd prime, then any quadratic form in at least three variables with at least three coefficients in $\mathbb{Z}_p^*$ is isotropic over $\mathbb{Q}_p$*. This is not only helpful in the proof of the theorem of Hasse-Minkowski but also a useful tool for concrete computations.

Exercise 41 *Prove that the equation*

$$3X^2 - 5Y^2 - 7Z^2 = 0$$

has a non-trivial solution in rational numbers x, y, z, while it fails to have a solution different from the trivial one when we change the sign by $5Y^2$.

We note some classic consequences of the theorem of Hasse-Minkowski:

- Lagrange proved that *every positive integer can be written as a sum of at most four squares*;
- Gauss showed that *every positive integer has a representation as a sum of at most three triangle numbers.*

For proofs of these corollaries see also [50]. The local-global principle has also limitations. For example, by the *theory of quadratic residues* it is easily seen that the equation

$$(X^2 - 2)(X^2 - 17)(X^2 - 34) = 0$$

has solutions in all $\mathbb{Q}_p, p \leq \infty$, but obviously it has no solution in $\mathbb{Q}$.

18 Hensel's lemma

Recall the construction of the 7-adic $\sqrt{2}$ from the last section. In a sense, solving the linear congruences step by step can be understood as applying the well-known *Newton iteration method* from real analysis to the polynomial $P(X) = X^2 - 2$. Starting with $x_1 = 3$ the iteration

$$x_n = x_{n-1} - \frac{P(x_{n-1})}{P'(x_{n-1})} = x_{n-1} - p\,\frac{P(x_n)}{p}\,(P'(x_{n-1}))^{-1},$$

yields

$$x_2 = 3 - 7\,\frac{3^2 - 2}{7}\,6^{-1} = 3 + 1 \cdot 7,$$

where $-6^{-1} = 1$ has to be understood as an equation in $\mathbb{Z}/7\mathbb{Z}$. Continuing this iteration process gives the same 7-adic value as we computed above. This observation leads to a very important theorem due to Hensel, called Hensel's lemma (which is also very useful in the proof of Theorem 36).

Theorem 37 *Let $P(X)$ be a polynomial with coefficients in $\mathbb{Z}_p$ and suppose that there is a p-adic integer x_1 such that*

$$P(x_1) \equiv 0 \bmod p\mathbb{Z}_p \qquad \textit{and} \qquad P'(x_1) \not\equiv 0 \bmod p\mathbb{Z}_p.$$

Then there exists a p-adic integer x with

$$x \equiv x_1 \bmod p\mathbb{Z}_p \qquad \textit{and} \qquad P(x) = 0.$$

Hensel's lemma is very likely the most important algebraic property of the p-adic numbers. In many circumstances one can decide very easily whether a polynomial has roots in $\mathbb{Z}_p$. The test involves finding an *approximation* x_1 of a root x and then verifying a condition on the (formal) derivative of the polynomial in question. This is certainly a highlight in combining diophantine approximations and diophantine equations! Moreover, where Newton's iteration method can fail (if the starting point of the iteration is not carefully chosen or the graph of the polynomial is not convex in a small neighbourhood of the root) Hensel's lemma always succeeds if the conditions are fulfilled.

Proof. The existence of the root x will follow from a construction of an appropriate Cauchy sequence converging to x. More precisely, we have to show that there are $x_n \in \mathbb{Z}_p$ satisfying

$$x_n \equiv x_{n+1} \bmod p^n \qquad \text{and} \qquad P(x_n) \equiv 0 \bmod p^n\mathbb{Z}_p.$$

The existence of x_1 follows from the assumption of the theorem. For x_2 we do the ansatz $x_2 = x_1 + zp$ (as in the proof of Theorem 34). In view of the formal identity (resp. Taylor expansion)

$$P(X+h) = P(X) + hP'(X) + \frac{h^2}{2!}P''(X) + \ldots,$$

we get

$$P(x_2) = P(x_1 + zp) \equiv P(x_1) + zpP'(x_1) \bmod p^2\mathbb{Z}_p.$$

Since $P(x_1) \equiv 0 \bmod p\mathbb{Z}_p$ there is some y for which $P(x_1) = yp$. Thus we have to solve the linear congruence

$$p\,(y + zP'(x_1)) \equiv 0 \bmod p^2\mathbb{Z}_p,$$

resp.

$$y + zP'(x_1) \equiv 0 \bmod p\mathbb{Z}_p.$$

Since $P'(x_1)$ is not divisble by p it is invertible in $\mathbb{Z}_p$. Thus we can take z with $0 \leq z < p$ and

$$z \equiv -y(P'(x_1))^{-1} \bmod p\mathbb{Z}_p.$$

This defines x_2 and by the same procedure we obtain the desired sequence by induction. It is clear that this sequence is Cauchy and, by continuity, the limit x satisfies $P(x) = 0$. Hensel's lemma is proved. •

A nice application of Hensel's lemma is to determine the roots of unity in $\mathbb{Q}_p$. Recall that ζ is called an n-th **root of unity** if $\zeta^n = 1$. We have to study $P(X) = X^n - 1$. For an n-th root of unity ζ we have $|\zeta^n|_p = 1$, and therefore $\zeta \in \mathbb{Z}_p^*$. It is easily seen that the roots of unity in $\mathbb{Q}_2$ are given by $\zeta = \pm 1$. The cases of odd primes is a bit more difficult. Suppose that $p \geq 3$ and let $x \in \mathbb{Z}_p$. Obviously, $P'(x) = nx^{n-1}$ is congruent zero modulo $p\mathbb{Z}_p$ if either p divides x, in which case x will not be a root of P anyway, or p divides n.

First, suppose that p and n are coprime. Then, by Hensel's lemma, for every $x_1 \not\equiv 0 \bmod p\mathbb{Z}_p$ we can find an $x \in \mathbb{Z}_p$ for which $P(x) = 0$ and $x \equiv x_1 \bmod p\mathbb{Z}_p$. It follows that the equation $P(X) = 0$ has as many distinct solutions in $\mathbb{Z}_p^*$ as in the group of prime residue classes. In view of *Fermat's little theorem* (59) $(\mathbb{Z}/p\mathbb{Z})^*$ consists exactly out of $p-1$ distinct $(p-1)$-th roots of unity. Consequently, each one corresponds to a $(p-1)$-th root of unity in $\mathbb{Q}_p$.

Now assume that n and p are not coprime. Without loss of generality we may suppose that $n = p$ (since each p-th root of unity is also a (kp)-th root of unity), and it remains to show that we cannot find any p-th root of unity $\neq 1$ in $\mathbb{Q}_p$. Let $x = x_1 + zp$, where $0 \leq x_1 < p$ and $z \in \mathbb{Z}_p$. Then

$$x^p - x_1^p = \sum_{k=1}^{p} \binom{p}{k} x_1^{p-k}(zp)^k,$$

which obviously lies in $p\mathbb{Z}_p$. If $x^p = 1$, then it follows from (59) that $x_1 = 1$. This leads to

$$1 = x^p = (1+zp)^p = 1 + zp^2 + \sum_{k=2}^{p-1} \binom{p}{k} (zp)^k + z^p p^p.$$

Suppose that $z \neq 0$, then

$$-zp^2 = \sum_{k=2}^{p-1} \binom{p}{k} (zp)^k + z^p p^p,$$

and, by (61),

$$|z|_p - 2 = |-zp^2|_p \leq \max_{2 \leq k \leq p-1} \left\{ \left| \binom{p}{k} z^k p^k \right|_p, |z^p p^p|_p \right\}.$$

A short computation gives a contradiction. So we have $z = 0$ which implies $x = 1$. Thus we have proved

Theorem 38 *Let p be an odd prime. $\mathbb{Q}_p$ contains exactly the $(p-1)$-th roots of unity.*

In particular, we see that any $\mathbb{Q}_p$ *is not algebraically closed.* (It would have been possible to see this before! Where?) This leads to new adventures (as in the real case) but this is another story...

You remember how he discovered the North Pole; well, he was so proud of this that he asked Christopher Robin if there were any other Poles such as a bear of little brain might discover. (A.A. Milne, *Winnie-the Pooh*)

References

[1] Archimedes, *The cattle problem*, in English verse by S.J.P. Hillion and H.W. Lenstra, Mercator, Santpoort 1999

[2] A. Baker, *Contributions to the theory of Diophantine equations II. The Diophantine equation $y^2 = x^3 + k$*, Phil. Trans. Roy. Soc. London **263** (1967/68), 193-208

[3] A. Baker, *Transcendental number theory*, Cambridge University Press 1975

[4] A. Baker, G. Wüstholz, *Number theory, transcendence and diophantine geometry in the next millenium*, 1-12, in 'Mathematics: frontiers and perspectives', V. Arnold et al. (editors), AMS 2000

[5] L. Berggren, J. Borwein, P. Borwein, *π: a scource book*, Springer 1997

[6] E. Bombieri, *Roth's theorem and the abc-conjecture*, preprint ETH Zürich, 1994

[7] D.M. Bressoud, *Factorization and primality testing*, Springer 1989

[8] P. Bundschuh, *Einführung in die Zahlentheorie*, Springer 1991, 2. Auflage

[9] E.B. Burger, *Exploring the number jungle: a journey into Diophantine Analysis*, AMS 2000

[10] H. Cohen, *A course in Computational Algebaric number theory*, Springer 1993

[11] J. Dieudonné, *Geschichte der Mathematik 1700-1900*, Vieweg 1978

[12] Ebbinghaus et al., *Zahlen*, Springer 1992, 3rd ed.

[13] N.D. Elkies, *ABC implies Mordell*, Intern. Math. Research Not. **7** (1991), 99-109

[14] C. Elsner, J.W. Sander, J. Steuding, *Kettenbrüche als Summen ebensolcher*, Math. Slov. **51** (2001), 281-293

[15] P. Erdös, *Representation of real numbers as sums and products of Liouville numbers*, Michigan Math. J. **9** (1962), 59-60

[16] C.J. Everett, *Fermat's conjecture, Roth's theorem, Pythagorian triangles, and Pell's equation*, Duke J. 801-804

[17] M. van Frankenhuysen, *The abc-conjecture implies Roth's theorem and Mordell's conjecture*, Math. Contemp. **16** (1999), 45-72

[18] F.Q. Gouvea, *p-adic numbers*, Springer 1993

[19] A. Granville, T.J. Tucker, *It's as easy as abc*, Notices A.M.S. (2002), 1224-1231

[20] M.Jr. Hall, *On the sum and product of continued fractions*, Ann. of Math. **28** (1947), 966-993

[21] G.H. Hardy, E.M. Wright, *An introduction to the theory of numbers*, Oxford University Press 1938

[22] K. Hensel, *Über eine neue Begründung der Theorie der algebraischen Zahlen*, Jahresberichte Deutschen Mathematiker Vereinigung **6** (1897), 83-88

[23] E. Hlawka, *Theorie der Gleichverteilung*, BIB 1979

[24] M.N. Huxley, *The distribution of prime numbers*, Oxford 1972

[25] K. Ireland, M. Rosen, *A classical introduction to modern number theory*, Springer 1990, 2nd ed.

[26] J.P. Jones, Y.V. Matijasevič, *Proof of recursive unsolvability of Hilbert's tenth problem*, Amer. Math. Monthly **98** (1991), 689-709

[27] A. Khintchine, *Kettenbrüche*, Teubner 1956

[28] N. Koblitz, *A course in Number theory and Cryptography*, Springer 1994, 2nd ed.

[29] S. Lang, *Introduction to diophantine approximations*, Springer 1995, 2n ed.

[30] S. Lang, *Complex Analysis*, Springer 1999, 4th ed.

[31] H.W. Lenstra, *Solving the Pell equation*, Notices Amer. Math. Soc. **49** (2002), 182-192

[32] A.A. Markoff, *Sur les formes binaires indefinies I+II*, Math. Ann. **15** (1879), 281-309; **17** (1880), 379-400

[33] R.C. Mason, *Diophantine equations over function fields*, LMS Lecture Notes **96**, Cambridge University Press 1984

[34] P. Mihăilescu, *Primary cyclotomic units and a proof for Catalan's conjecture*, preprint

[35] L.J. Mordell, *Diophantine equations*, Academic Press 1969

[36] M.B. Nathanson, *Polynomial Pell equations*, Proc. A.M.S. **56** (1976), 89-92

[37] I. Niven, *A simple proof that π is irrational*, Bull. Amer. Math. Soc. **53** (1947), 509

[38] O. Perron, *Die Lehre von den Kettenbrüchen*, vol. I, Teubner 1954

[39] A. van der Poorten, *A proof that Euler missed*, Math. Intelligencer 195-203

[40] P. Ribenboim, *The book of prime number records*, Springer 1989, 2nd ed.

[41] G.J. Rieger, *Zahlentheorie*, Vandenhoeck & Ruprecht, Göttingen 1976

[42] K.F. Roth, *Rational approximations to algebraic numbers*, Proceedings of the ICM 1958, Edinburgh, Cambridge University Press, 203-210

[43] H.P. Schlickewei, *The subspace theorem and applications*, Doc. Math. J. DMV, Extra volume ICM 1998, vol. II, 197-206

[44] W.M. Schmidt, *Diophantine Approximation*, Springer 1980, Lecture Notes 785

[45] T. Schneider, *Einführung in die transzendenten Zahlen*, Springer 1957

[46] M.R. Schroeder, *Number theory in science and communications*, Springer 1997, 3rd ed.

[47] F. Schwarz, *Wieviele Rinder hat der Sonnengott?*, Mitteilungen der Deutschen Mathematiker Vereinigung **2** (1997), 13-18

[48] W. Schwarz, *Geschichte der Zahlentheorie*, www.math.uni-frankfurt.de/ steuding/schwarz.html

[49] C. Series, *The geometry of Markoff numbers*, Math. Intelligencer **7**, no.3 (1985), 20-29

[50] J.P. Serre, *A course in Arithmetic*, Springer 1973

[51] J.H. Silverman, J. Tate, *Rational points on elliptic curves*, Springer 1992

[52] S. Singh, *Fermat's last theorem*, Fourth Estate, London 1997

[53] R. Šleževicienė, J. Steuding, *Simpson's paradox in the Farey sequence*, preprint

[54] R. Šleževicienė, J. Steuding, *Factoring with continued fractions and the Pell equation*, in preparation

[55] J. Steuding, *The world of p-adic numbers and p-adic functions*, Proc. Sci. Seminar Fac. Phys. Math. Šiauliai Univ. **5** (2002), 90-107

[56] J. Steuding, *A note on the Polynomial Pell equation*, preprint 2003

[57] C.L. Stewart, Kunrui Yu, *On the abc conjecture, II*, Duke J. **108** (2001), 169-

[58] W.W. Stothers, *Polynomial identities and hauptmoduln*, Quart. J. Math. **32** (1981), 349-370

[59] P. Vojta, *Nevanlinna theory and diophantine approximation*, Several complex variables, Berkely, CA, 1995/96, Math. Sci. Res. Inst. Publ. **37**, Cambridge University Press 1999

[60] G. Wüstholz, *Ausgewählte Kapitel aus der Zahlentheorie und Geometrie*, Lecture Notes 1996, www.math.uga.edu/ ntheory/N4.html

[61] W. Zudilin, *Arithmetic of linear forms involving odd zeta values*, (2001), arXiv:math.NT