INTRODUCTION TO NUMBER THEORY

MARTIN KLAZAR

(lecture notes)

These lecture notes cover the one-semester course Introduction to Number Theory (Úvod do teorie čísel, MAI040) that I have been teaching on the Faculty of Mathematics and Physics of Charles University in Prague since 1996. Needless to say, I do not claim any originality of the material presented here (in most cases I attempted to give references to the corresponding secondary sources that I used).

July 2006 Martin Klazar

Contents

1 Diophantine approximation **1**
- 1.1 Dirichlet's theorem 1
- 1.2 Farey fractions and a theorem of Hurwitz 4
- 1.3 Continued fractions 7
- 1.4 Transcendental numbers 12
- 1.5 Remarks 18

2 Diophantine equations **19**
- 2.1 Three famous exproblems 19
- 2.2 Pell equation 21
- 2.3 Thue equation 24
- 2.4 FLT for $n = 2, 4$ and for polynomials 26
- 2.5 Remarks 31

3 Geometry of numbers **32**
- 3.1 Lattices, Farey fractions and convex bodies 32
- 3.2 Four-squares theorem 36
- 3.3 The circle problem and the divisor problem 40
- 3.4 Remarks 43

4 Prime numbers **44**
- 4.1 Four proofs of the infinititude of primes 44
- 4.2 Theorems of Chebyshev and Mertens 47
- 4.3 Estimates of the functions $\omega(n)$ and $\Omega(n)$ 54
- 4.4 Remarks 60

5 Congruences **61**
5.1 Quadratic residues and the quadratic reciprocity law 61
5.2 Remarks 67

6 Integer partitions **68**
6.1 Compositions and partitions 68
6.2 Euler's identity and a metaidentity of Cohen and Remmel . . 71
6.3 The pentagonal identity 77
6.4 Remarks 82

References **83**

Index of names **85**

Chapter 1

Diophantine approximation

This discipline of number theory investigates to what extent real numbers can be approximated by fractions. We prove Dirichlet's theorem which says that every irrational number can be approximated by infinitely many fractions p/q with precision better than q^{-2}. A nice application is that every prime number of the form $4n + 1$ is a sum of two squares. We introduce Farey fractions and prove by means of them a result of Hurwitz that gives a best possible strengthening of Dirichlet's theorem. An important tool in approximation (of not only numbers) is continued fractions. We develop some of their basic properties in section 1.3. We present the argument of Liouville producing transcendental (i.e., non-algebraic) numbers and give Hilbert's proof of the transcendence of Euler's number e.

1.1 Dirichlet's theorem

Let us review some notation. $\mathbf{N} = \{1, 2, \ldots\}$ are natural numbers, $\mathbf{N}_0 = \{0, 1, 2, \ldots\}$, $\mathbf{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ are integers, $\mathbf{Q} = \{a/b : \ a \in \mathbf{Z}, b \in \mathbf{N}\}$ are rational numbers (fractions), $\mathbf{R}$ are real numbers, and $\mathbf{C} = \{a + bi : \ a, b \in \mathbf{R}\}$ are complex numbers ($i = \sqrt{-1}$). For $\alpha \in \mathbf{R}$, the *integer part of* α and the *fractional part of* α are, respectively,

$$\lfloor \alpha \rfloor = \max\{m \in \mathbf{Z} : \ m \le \alpha\} \ \text{ and } \ \{\alpha\} = \alpha - \lfloor \alpha \rfloor \in [0, 1).$$

Another notation is $\lceil\alpha\rceil = \min\{m \in \mathbf{Z} : \ m \geq \alpha\}$ and $\|\alpha\| = \min(\{\alpha\}, 1 - \{\alpha\})$ (distance to the nearest integer).

The following fundamental theorem in diophantine approximation is due to Peter Dirichlet (1805–1859).

Theorem (Dirichlet, 1842).

1. *For every $\alpha \in \mathbf{R}$ and $Q \in \mathbf{N}$, $Q \geq 2$, there exist numbers $p, q \in \mathbf{Z}$ such that $1 \leq q < Q$ and*

$$\left|\alpha - \frac{p}{q}\right| \leq \frac{1}{qQ}.$$

2. *For every irrational $\alpha \in \mathbf{R}$ the inequality*

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

has infinitely many rational solutions p/q.

Proof. 1. We split $[0,1]$ in Q subintervals $[(i-1)/Q, i/Q]$, $i = 1, 2, \ldots, Q$, each with length $1/Q$, and consider $Q+1$ numbers $0, 1, \{\alpha\}, \{2\alpha\}, \ldots \{(Q-1)\alpha\}$. They lie in $[0,1]$ and each of them has form $s\alpha - r$ for some $r, s \in \mathbf{Z}$ with $0 \leq s < Q$ ($0 = 0\alpha - 0$, $1 = 0\alpha - (-1)$, and $\{i\alpha\} = i\alpha - \lfloor i\alpha \rfloor$ for $i = 1, 2, \ldots, Q-1$). By the pigeonhole principle, two numbers must fall in the same subinterval. Note that these two numbers cannot be 0 and 1. Hence

$$|(s_1\alpha - r_1) - (s_2\alpha - r_2)| \leq \frac{1}{Q}$$

for some $r_i, s_i \in \mathbf{Z}$ with $0 \leq s_2 < s_1 < Q$. Setting $q = s_1 - s_2$ and $p = r_1 - r_2$, we get the desired fraction p/q because $1 \leq q < Q$ and

$$|q\alpha - p| = |(s_1 - s_2)\alpha - (r_1 - r_2)| = |(s_1\alpha - r_1) - (s_2\alpha - r_2)| \leq \frac{1}{Q}.$$

2. Suppose we have already solutions $p_1/q_1, \ldots, p_r/q_r$. We select $Q \in \mathbf{N}$ so big that

$$\frac{1}{Q} < \Delta = \min_{i=1\ldots r} \left|\alpha - \frac{p_i}{q_i}\right|.$$

(If we have no solution yet, we select Q arbitrarily. Since α is irrational, certainly $\Delta > 0$.) By part 1 there is a fraction $p/q \in \mathbf{Q}$ such that $1 \leq q < Q$ and

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

Also, $1/qQ \le 1/Q < \Delta$. Thus $p/q \ne p_i/q_i$ for every $i = 1, 2, \ldots, r$ and p/q is a new solution. This way we obtain infinitely many solutions. $\square$

As an application of Dirichlet's theorem we prove a nice result that was stated first by Pierre de Fermat (1601–1665) and proved later by Leonhard Euler (1707–1783).

Theorem (Euler, 1747). *Every prime number p of the form $4n+1$ is a sum of two squares, $p = a^2 + b^2$ for some $a, b \in \mathbf{N}$.*

For example, we have representations $5 = 2^2+1^2$, $13 = 3^2+2^2$, $17 = 4^2+1^2$, $29 = 5^2+2^2$, $37 = 6^2+1^2$, and $41 = 5^2+4^2$. The primes of the form $4n+3$ are never sum of two squares because squares give modulo 4 only residues 0 and 1. We need a lemma which belongs to the theory of quadratic residues (see chapter 5).

Lemma. *For every prime p of the form $4n+1$ there is a $c \in \mathbf{N}$ such that $c^2 \equiv -1 \pmod{p}$.*

Proof of the theorem. For a given prime $p = 4n+1$ we take a $c \in \mathbf{N}$ satisfying $c^2 \equiv -1 \pmod{p}$ and set $\alpha = c/p$, $Q = \lceil \sqrt{p} \rceil$. By part 1 of Dirichlet's theorem, there are numbers $a, b \in \mathbf{Z}$ satisfying $1 \le b < \sqrt{p}$ and

$$\left| \frac{c}{p} - \frac{a}{b} \right| < \frac{1}{b\sqrt{p}}.$$

Thus $0 \le |cb - pa| < \sqrt{p}$ and $0 < (cb-pa)^2 + b^2 < 2p$. But the number $(cb-pa)^2 + b^2 = (c^2+1)b^2 + p(pa^2 - 2cba)$ is divisible by p (due to the selection of c) and therefore $(cb-pa)^2 + b^2 = p$. $\square$

Proof of the lemma. We consider the finite field $\mathbf{Z}_p$ of residues modulo p and its subsets

$$M_x = \{x, -x, x^{-1}, -x^{-1}\} \quad \text{where} \quad x \in \mathbf{Z}_p^* = \mathbf{Z}_p \backslash \{0\}.$$

They form a set partition of $\mathbf{Z}_p^*$ because $x \in M_x$, $0 \in M_x$ for no x, and $M_x \cap M_y \ne \emptyset$ implies that $M_x = M_y$. For example, $\mathbf{Z}_{13}^* = \{1, 2, \ldots, 12\}$ is partitioned in the sets $\{2, 11, 7, 6\}$, $\{3, 10, 9, 4\}$, $\{1, 12\}$, and $\{5, 8\}$. Since $x \ne -x$ $(p > 2)$, we have $|M_x| = 4$ or $|M_x| = 2$. The latter happens only if $x = x^{-1}$ or $x = -x^{-1}$. Because $x = x^{-1}$ is equivalent with $(x+1)(x-1) = 0$,

$x = x^{-1}$ produces the set $M_1 = M_{-1} = \{1, -1\}$. The second case $x = -x^{-1}$ is equivalent with $x^2 = -1$. But $|\mathbf{Z}_p^*| = p - 1 = 4n$ and therefore $\mathbf{Z}_p^*$ cannot be partitioned in four-element sets and one two-element set. There must be another two-element set M_x, which means that there is an element x in the field $\mathbf{Z}_p$ satisfying $x^2 = -1$. $\square$

1.2 Farey fractions and a theorem of Hurwitz

Another way how to prove Dirichlet's theorem is via *Farey fractions*. For a given $n \in \mathbf{N}$, these are the fractions $p/q \in [0, 1]$ in lowest terms which have denominator $q \leq n$. Sorted by size, they form the list $\mathcal{F}_n$ of *Farey fractions of order* n. For example,

$$\mathcal{F}_5 = \left(\frac{0}{1}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{2}{5}, \frac{1}{2}, \frac{3}{5}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \frac{1}{1}\right).$$

For any two distinct fractions $a/b < c/d$ we have $c/d - a/b \geq 1/bd$. If equality occurs, we say that a/b and c/d are *as close as possible*. Interestingly, two consecutive Farey fractions in $\mathcal{F}_n$ are always as close as possible! This was stated as a problem by geologist John Farey (1766–1826) in 1816 and proved immediately afterwards by Augustin Cauchy (1789–1857). (In fact, Farey stated that if $a_1/b_1 < a_2/b_2 < a_3/b_3$ are three consecutive members of $\mathcal{F}_n$, then $a_2/b_2 = (a_1 + a_3)/(b_1 + b_3)$. This is equivalent to the property that two consecutive members are as close as possible.)

Theorem (Farey–Cauchy, 1816). *If $a/b < c/d$ are two consecutive members of $\mathcal{F}_n$ then $bc - ad = 1$, which means that a/b and c/d are as close as possible.*

Proof. We consider the diophantine equation

$$bx - ay = 1$$

with unknowns $x, y \in \mathbf{Z}$. Since a and b are coprime, it has at least one solution. (Division algorithm shows that the ideal $\{bx + ay : \ x, y \in \mathbf{Z}\}$ in the ring of integers is generated by the greatest common divisor of a and b, which is 1. So 1 is an integral linear combination of a and b.) Our aim is to show that c, d is a solution.

If x, y is a solution, so is $x - ra, y - rb$ for any $r \in \mathbf{Z}$. Hence there is a solution $x_1, y_1 \in \mathbf{Z}$ such that

$$n - b < y_1 \leq n.$$

Rearranging $bx_1 - ay_1 = 1$, we express x_1/y_1 as

$$\frac{x_1}{y_1} = \frac{1}{by_1} + \frac{a}{b}.$$

We have $x_1/y_1 \in \mathcal{F}_n$. (Numbers x_1 and y_1 are coprime by $bx_1 - ay_1 = 1$. We know that $0 < y_1 \leq n$. Thus, by $bx_1 - ay_1 = 1$ and $0 < a < b$, we have $0 < x_1 < y_1$.) It follows that $x_1/y_1 \geq c/d$. We show that $x_1/y_1 > c/d$ leads to a contradiction.

Let $x_1/y_1 > c/d$. Adding the trivial inequalities

$$\frac{x_1}{y_1} - \frac{c}{d} \geq \frac{1}{dy_1} \quad \text{and} \quad \frac{c}{d} - \frac{a}{b} \geq \frac{1}{bd}$$

we get

$$\left(\frac{1}{by_1} =\right) \ \frac{x_1}{y_1} - \frac{a}{b} \geq \frac{1}{dy_1} + \frac{1}{bd} = \frac{b + y_1}{bdy_1}$$

where the equality in brackets follows from the above expression for x_1/y_1. Multiplying by bdy_1, we get the inequality $d \geq b + y_1$. But $b + y_1 > n$ by the above bound on y_1. So $d > n$, which contradicts $c/d \in \mathcal{F}_n$.

Therefore $x_1/y_1 = c/d$. Since these are fractions in lowest terms, $x_1 = c$ and $y_1 = d$. We have proved that c, d is a solution of $bx - ay = 1$. □

In chapter 3 we give a geometric proof. This theorem leads to another proof of part 2 of Dirichlet's theorem. Suppose $\alpha \in (0, 1)$ is irrational. We squeeze α between two consecutive elements of $\mathcal{F}_n$: $a/b < \alpha < c/d$. Let $b \leq d$ (for $b > d$ we proceed similarly). Then

$$0 < \alpha - \frac{a}{b} < \frac{c}{d} - \frac{a}{b} = \frac{1}{bd} \leq \frac{1}{b^2}$$

where the equality is provided by the Farey–Cauchy theorem. To obtain another approximation, we use the fact that the distance between two consecutive members of $\mathcal{F}_m$ is at most $1/m$. Let m be so big that $1/m < \min(\alpha - a/b, c/d - \alpha)$. Then if we squeeze α between two consecutive elements of $\mathcal{F}_m$, $e/f < \alpha < g/h$, we have $e/f \neq a/b$, $g/h \neq c/d$ and e/f or g/h is a new close approximation. Continuing this way, we obtain infinitely many fractions p/q satisfying $|\alpha - p/q| < 1/q^2$.

We are coming to a result of Adolf Hurwitz (1859–1919) that characterizes the best Dirichlet-type inequality.

Theorem (Hurwitz, 1891). 1. *For every irrational $\alpha \in \mathbf{R}$ the inequality*

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

has infinitely many rational solutions p/q.

2. *For every real $A > \sqrt{5}$ the inequality*

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

has only finitely many rational solutions p/q.

Proof. 1. We may suppose that $\alpha \in (0,1)$. We show that if $a/b < \alpha < c/d$ for two consecutive Farey fractions from $\mathcal{F}_n$, then one of the three fractions

$$\frac{a}{b},\ \frac{c}{d},\ \frac{e}{f} = \frac{a+c}{b+d}$$

satisfies the inequality. Squeezing α between two consecutive members of $\mathcal{F}_n$ for larger and larger n, we obtain (as for for Dirichlet's inequality) infinitely many fractions satisfying $|\alpha - p/q| < 1/\sqrt{5}q^2$.

Let us assume for the contradiction that neither of the three fractions satisfies the inequality. We may suppose that $\alpha > e/f$, for $\alpha < e/f$ we proceed similarly. Thus

$$\alpha - \frac{a}{b} \geq \frac{1}{\sqrt{5}b^2},\ \alpha - \frac{e}{f} \geq \frac{1}{\sqrt{5}f^2},\ \frac{c}{d} - \alpha \geq \frac{1}{\sqrt{5}d^2}.$$

Note that equalities may occur. Adding the first and the third inequality, and the second and the third inequality, we get

$$\frac{1}{bd} = \frac{c}{d} - \frac{a}{b} \geq \frac{1}{\sqrt{5}}\left(\frac{1}{b^2} + \frac{1}{d^2}\right) \quad \text{and} \quad \frac{1}{df} = \frac{c}{d} - \frac{e}{f} \geq \frac{1}{\sqrt{5}}\left(\frac{1}{f^2} + \frac{1}{d^2}\right),$$

where the equalities follow from the theorem on Farey fractions. Multiplying the first inequality by $\sqrt{5}b^2d^2$, the second one by $\sqrt{5}d^2f^2$, and adding the results, we obtain

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

which is equivalent with

$$0 \geq \tfrac{1}{2}((\sqrt{5}-1)d - 2b)^2.$$

This implies $(\sqrt{5}-1)d - 2b = 0$ and $\sqrt{5} = 1 + 2b/d \in \mathbf{Q}$, which is a contradiction.

2. We denote $\beta = (\sqrt{5}-1)/2$. We fix $A > \sqrt{5}$ and suppose that $|\beta - p/q| < 1/Aq^2$ for infinitely many fractions p/q. Hence q may be as large as we wish. In other words,

$$\beta = \frac{p}{q} + \frac{\delta}{q^2}$$

has infinitely many solution $p/q, \delta$ where $p/q \in \mathbf{Q}$ and $\delta \in \mathbf{R}$, $|\delta| < 1/A$. We rewrite this as

$$\frac{\delta}{q} - \frac{q\sqrt{5}}{2} = q\beta - p - \frac{q\sqrt{5}}{2} = -\frac{q}{2} - p.$$

Squaring and subtracting $5q^2/4$, we get the identity

$$\frac{\delta^2}{q^2} - \delta\sqrt{5} = p^2 + pq - q^2.$$

For sufficiently big q the left side is in absolute value smaller than 1 because $\delta^2/q^2 \to 0$ as $q \to \infty$ and $|\delta\sqrt{5}| < \sqrt{5}/A < 1$. This means that $p^2+pq-q^2 = 0$ has a solution $p, q \in \mathbf{Z}$ ($|z| < 1$ for $z \in \mathbf{Z}$ means that $z = 0$). But the last equation is equivalent with $(2p+q)^2 = 5q^2$, which gives again the contradiction $\sqrt{5} = 1 + 2p/q \in \mathbf{Q}$. □

1.3 Continued fractions

We have seen two approaches to Dirichlet-type approximations of irrational numbers, the pigeonhole principle and Farey fractions. Now we demonstrate the third and most important approach, that of continued fractions.

To approximate a real number $\alpha \in \mathbf{R}$ by fractions we approximate it first by $a_0 = \lfloor \alpha \rfloor \in \mathbf{Z}$. If $\zeta_0 = \alpha - a_0 = \{\alpha\} \in [0,1)$ is zero, we finish the procedure. If $\zeta_0 \neq 0$, we write $1/\zeta_0 = a_1 + \zeta_1$ where $a_1 = \lfloor 1/\zeta_0 \rfloor \in \mathbf{N}$ and $\zeta_1 = \{1/\zeta_0\} \in [0,1)$. If $\zeta_1 = 0$, we finish the procedure. If $\zeta_1 \neq 0$, we express again $1/\zeta_1 = a_2 + \zeta_2$ with $a_2 \in \mathbf{N}$ and $\zeta_2 \in [0,1)$. Continuing, in the n-th step we get the expression

$$\alpha = a_0 + \zeta_0 = a_0 + \frac{1}{a_1 + \zeta_1} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \zeta_2}} = \cdots$$

$$= \quad a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\begin{matrix} \ddots \\ a_{n-1} + \cfrac{1}{a_n + \zeta_n} \end{matrix}}}}$$

where $a_0 \in \mathbf{Z}$, $a_i \in \mathbf{N}$ for $i > 0$, and $\zeta_i \in [0, 1)$. We may hope to get a good rational approximation of α by replacing ζ_n with zero. We show shortly that this hope is justified.

Sequence $[a_0, a_1, a_2, \ldots]$ is the *continued fraction (expansion) of* α. Numbers a_i are the *terms* of the continued fraction and the fractions obtained by replacing the ζ_i's with 0 are called *convergents* of α. Let us look at two examples of continued fractions.

$$\begin{aligned} -\frac{119}{27} &= -5 + \frac{16}{27} = -5 + \cfrac{1}{1 + \cfrac{11}{16}} = -5 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{5}{11}}} \\ &= -5 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{5}}}} = [-5, 1, 1, 2, 5]. \end{aligned}$$

Convergents of $-119/27$ are $-\frac{5}{1}, -\frac{4}{1}, -\frac{9}{2}, -\frac{22}{5}$, and $-\frac{119}{27}$. The golden ratio $\phi = (1 + \sqrt{5})/2 = 1.61803\ldots$ has a simple continued fraction. Since

$$\frac{1 + \sqrt{5}}{2} = 1 + \frac{\sqrt{5} - 1}{2}$$

with $(\sqrt{5} - 1)/2 \in [0, 1)$, we have $a_0 = 1, \zeta_0 = (\sqrt{5} - 1)/2$. In the second step,

$$\frac{1}{\zeta_0} = \frac{2}{\sqrt{5} - 1} = \frac{1 + \sqrt{5}}{2}$$

and we are back at the beginning. Thus

$$\frac{1 + \sqrt{5}}{2} = [1, 1, 1, 1, \ldots].$$

It is convenient to view convergents more generally as rational functions:

$$\begin{aligned}[x_0, x_1, \dots, x_n] &= \frac{p_n}{q_n} = \frac{p_n(x_0, x_1, \dots, x_n)}{q_n(x_0, x_1, \dots, x_n)} \\ &= x_0 + \cfrac{1}{x_1 + \cfrac{1}{x_2 + \cfrac{1}{\ddots \\ x_{n-1} + \cfrac{1}{x_n}}}}\end{aligned}$$

where x_i are variables and p_n, q_n are coprime polynomials. If $p'_n/q'_n \in \mathbf{Q}$ is the n-th convergent of $\alpha \in \mathbf{R}$ with continued fraction $[a_0, a_1, \dots]$, then $p'_n = p_n(a_0, \dots, a_n)$ and $q'_n = q_n(a_0, \dots, a_n)$.

Numerators and denominators of convergents can be calculated by simple recurrences. In fact, it is a single recurrence, only with different initial conditions for denominators and for numerators.

Lemma. *We have the recurrence $p_0 = x_0$, $p_1 = x_0x_1 + 1$, $q_0 = 1$, $q_1 = x_1$, and, for $n \geq 2$,*

$$p_n = x_n p_{n-1} + p_{n-2} \quad \textit{and} \quad q_n = x_n q_{n-1} + q_{n-2}.$$

Proof. By induction on n. For $n = 0, 1$ the lemma holds. Suppose it holds for the n-th convergent. For the $(n+1)$-th one we have

$$\begin{aligned}\frac{p_{n+1}}{q_{n+1}} &= [x_0, \dots, x_{n-1}, x_n, x_{n+1}] = [x_0, \dots, x_{n-1}, x_n + 1/x_{n+1}] \\ &= \frac{(x_n + 1/x_{n+1})p_{n-1} + p_{n-2}}{(x_n + 1/x_{n+1})q_{n-1} + q_{n-2}} \\ &= \frac{x_{n+1}(x_n p_{n-1} + p_{n-2}) + p_{n-1}}{x_{n+1}(x_n q_{n-1} + q_{n-2}) + q_{n-1}} \\ &= \frac{x_{n+1}p_n + p_{n-1}}{x_{n+1}q_n + q_{n-1}}\end{aligned}$$

where on the second and fourth line we used the inductive assumption. □

Strictly speaking, we have shown that the recurrence calculates convergents but it may not be clear that it produces convergents in lowest terms. However, if p_n and q_n are calculated by the recurrence, we have

$$\begin{aligned}
p_n q_{n-1} - q_n p_{n-1} &= (x_n p_{n-1} + p_{n-2}) q_{n-1} - (x_n q_{n-1} + q_{n-2}) p_{n-1} \\
&= -(p_{n-1} q_{n-2} - q_{n-1} q_{n-2}) \\
&\vdots \\
&= (-1)^{n-1} (p_1 q_0 - q_1 p_0) \\
&= (-1)^{n-1},
\end{aligned}$$

which shows that p_n, q_n are indeed coprime. By the same argument this is true also for numbers—the convergents $p_n/q_n \in \mathbf{Q}$ of $\alpha \in \mathbf{R}$ calculated by the recurrence from the continued fraction $[a_0, a_1, \ldots]$ of α are in lowest terms. Using the recurrence and the last identity we get

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

We rewrite both identities as follows.

Lemma. *For every n, for which the convergents are defined, we have*

$$\frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}} = \frac{(-1)^{n-1}}{q_{n-1} q_n} \quad \textit{and} \quad \frac{p_n}{q_n} - \frac{p_{n-2}}{q_{n-2}} = \frac{(-1)^n x_n}{q_{n-2} q_n} \ .$$

In the next theorem we summarize basic properties of continued fractions. In particular, we give yet another proof of the second part of Dirichlet's theorem.

Theorem. *Let $p_n/q_n \in \mathbf{Q}$ be the n-th convergent of the continued fraction expansion $[a_0, a_1, \ldots]$ of $\alpha \in \mathbf{R}$.*

1. *For every $n \geq 0$,*

$$\frac{p_{2n}}{q_{2n}} \leq \alpha \leq \frac{p_{2n+1}}{q_{2n+1}}.$$

2. *The continued fraction $[a_0, a_1, \ldots]$ is finite iff α is rational.*

3. *Rational α equals to its last convergent, $\alpha = p_n/q_n = [a_0, a_1, \ldots, a_n]$.*

4. *For irrational α we have the inequalities*

$$\frac{p_0}{q_0} < \frac{p_2}{q_2} < \frac{p_4}{q_4} < \cdots < \alpha < \cdots < \frac{p_5}{q_5} < \frac{p_3}{q_3} < \frac{p_1}{q_1}$$

and $p_n/q_n \to \alpha$ as $n \to \infty$.

5. *For every $n \in \mathbf{N}$,*

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

Proof. 1. This follows from the expression of α in terms of a_i and ζ_i we started with. If $\zeta_n \neq 0$ is replaced with 0, the compounded fraction decreases for even n and increases for odd n. 2. Since for irrational α never $\zeta_n = 0$, the procedure never terminates and $[a_0, a_1, \ldots]$ is an infinite sequence. For rational $\alpha = p/q$ it is not hard to see that the procedure generating $[a_0, a_1, \ldots]$ is in fact a reformulation of Euclid's algorithm determining the gcd of p, q and therefore must terminate. 3. Clear because $\zeta_n = 0$ in the last step. 4. This follows from the previous lemma and from part 1. 5. This follows from part 4, the first identity of the previous lemma, and from $q_n < q_{n+1}$ (which follows from the recurrence for the q_n's). $\square$

In 1737, L. Euler discovered that the continued fraction of "his" number $\mathrm{e} = 2.71828\ldots$ follows a nice simple rule:

$$\mathrm{e} = [2, 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, 1, 1, 10, 1, 1, 12, 1, 1, 14, \ldots].$$

In contrast, the continued fraction of $\pi = 3.14159\ldots$ does not follow any apparent rule (and no such rule is known):

$$\pi = [3, 7, 15, 1, 292, 1, 1, 1, 2, 1, 3, 1, 14, 2, 1, 1, 2, 2, 2, 2, 1, 84, 2, 1, 1, 15, 3, \ldots].$$

We say that an infinite sequence $a_0, a_1, \ldots$ is *eventually periodic* if there exist integers $r \geq 0, s \geq 1$ such that $a_i = a_{i+s}$ for every $i \geq r$. We shall not prove the result of Joseph-Louis Lagrange (1736–1813) characterizing eventually periodic continued fractions.

Theorem (Lagrange, 1770). *Let $\alpha \in \mathbf{R}$ be irrational. The continued fraction of α is eventually periodic if and only if $a\alpha^2 + b\alpha + c = 0$ for some integers a, b, c, not all zero.*

Using continued fractions, one can characterize the "most irrational" numbers $\alpha \in \mathbf{R}$. Let $\alpha = [a_0, a_1, \ldots]$ be irrational. If $a_i = 1$ for all $i \geq i_0$, then for every $A > \sqrt{5}$ the inequality

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

has only finitely many solutions $p/q \in \mathbf{Q}$. If $a_i \geq 2$ for infinitely many i, then the inequality

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{\sqrt{8}q^2}$$

has infinitely many solutions $p/q \in \mathbf{Q}$. This was proved by Hurwitz.

1.4 Transcendental numbers

Recall that $\alpha \in \mathbf{R}$ is *algebraic* if it is a root of a nonzero integral polynomial, that is,

$$a_n\alpha^n + a_{n-1}\alpha^{n-1} + \cdots + a_1\alpha + a_0 = 0$$

for some $a_i \in \mathbf{Z}$, not all of them zero. The smallest degree n of such a polynomial is the *degree* of α. Rational numbers are exactly the algebraic numbers with degree 1. Numbers, which are not algebraic, are called *transcendental*. Joseph Liouville (1809–1882) proved that algebraic numbers do not have too close rational approximations.

Theorem (Liouville, 1844). *Let $\alpha \in \mathbf{R}$ be an algebraic number with degree $n \geq 2$ (i.e., α is irrational). There exists a constant $c = c(\alpha) > 0$ depending only on α such that*

$$\left|\alpha - \frac{p}{q}\right| \geq \frac{c}{q^n}$$

for every $p/q \in \mathbf{Q}$.

Proof. Let $P(x) = a_nx^n + \cdots + a_1x + a_0$ be nonzero integral polynomial with root α and with the lowest degree. We set $I = [\alpha - 1, \alpha + 1]$ and

$$c = \min(1, (\max_{x \in I} |P'(x)|)^{-1}).$$

If $p/q \notin I$, the inequality holds trivially:

$$\left|\alpha - \frac{p}{q}\right| \geq 1 \geq \frac{1}{q^n} \geq \frac{c}{q^n}.$$

If $p/q \in I$, Lagrange's mean value theorem asserts that

$$\frac{P(\alpha) - P(p/q)}{\alpha - p/q} = P'(z)$$

for some real number z lying between α and p/q and hence in I. Since $P(\alpha) = 0$, we obtain

$$\left|\alpha - \frac{p}{q}\right| = \frac{|P(p/q)|}{|P'(z)|}.$$

By the definition of c, $1/|P'(z)| \geq c$. Also, $P(p/q) \neq 0$ because otherwise $P(x)/(x-p/q)$ would be a rational polynomial with root α and degree $n-1$, contradicting the minimality of n. But then

$$|P(p/q)| = \frac{|a_n p^n + \cdots + a_1 p q^{n-1} + a_0 q^n|}{q^n} \geq \frac{1}{q^n}$$

because the numerator is a nonzero integer. Thus

$$\left|\alpha - \frac{p}{q}\right| \geq \frac{c}{q^n}.$$

□

Liouville's theorem provides a method for producing transcendental numbers.

Corollary. *Let $\alpha \in \mathbf{R}$ be an irrational number with the property that for every $n \in \mathbf{N}$ there is a fraction p/q satisfying $q \geq 2$ and*

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

Then α is a transcendental number.

Proof. Suppose, for the contradiction, that α is algebraic and has degree $m \geq 2$. Let $c = c(\alpha)$ be the constant from Liouville's theorem. We select $n \in \mathbf{N}$ big enough so that $n > m$ and $2^{m-n} < c$. By the property of α, there exists a fraction p/q that has $q \geq 2$ and is closer to α than $1/q^n$. But then

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{q^n} \leq \frac{1}{q^m} \cdot \frac{1}{2^{n-m}} < \frac{c}{q^m},$$

which contradicts Liouville's theorem. □

Real numbers with the property stated in the corollary are called *Liouville numbers*. Corollary says that every Liouville number is transcendental. Liouville numbers can be constructed as sums of rapidly converging infinite series; the required close rational approximations are provided by partial sums. In this way one easily shows that

$$\sum_{n=1}^{\infty} \frac{1}{10^{i!}} = 0.110001000000000000000001000000000000000000000000000\ldots$$

is Liouville number and hence it is transcendental.

Norwegian mathematician Axel Thue (1863-1922) obtained an important strengthening of Liouville's theorem.

Theorem (Thue, 1909). *Let $\alpha \in \mathbf{R}$ be an algebraic number with degree $n \geq 2$ and let $\varepsilon > 0$. There exists a constant $c = c(\alpha, \varepsilon) > 0$ such that*

$$\left|\alpha - \frac{p}{q}\right| \geq \frac{c}{q^{1+\varepsilon+n/2}}$$

holds for every $p/q \in \mathbf{Q}$.

For $n = 2$ the lower bound is in fact weaker (i.e., it is asymptotically smaller) than in Liouville's theorem but for $n \geq 3$ it is much stronger. Thue's theorem has important consequences in the theory of diophantine equations, as we will see in the next chapter. It inspired further strengthenings which replaced exponent $1 + \varepsilon + n/2$ by smaller functions of n. I will not mention these intermediate improvements obtained by Siegel, Gelfond, and Dyson and I state only the last ultimate result due to Klaus Roth (1925), for which he was in 1958 awarded the Fields medal.

Theorem (Roth, 1955). *Let $\alpha \in \mathbf{R}$ be an algebraic number with degree $n \geq 2$ and let $\varepsilon > 0$. There exists a constant $c = c(\alpha, \varepsilon) > 0$ such that*

$$\left|\alpha - \frac{p}{q}\right| \geq \frac{c}{q^{2+\varepsilon}}$$

for every $p/q \in \mathbf{Q}$.

An equivalent formulation of Roth's theorem (and similarly for Thue's theorem) is: *For every irrational algebraic number α and every $\varepsilon > 0$ the inequality $|\alpha - p/q| < 1/q^{2+\varepsilon}$ has only finitely many rational solutions p/q.*

Let us recall the proof of the irrationality of Euler's number e. For the contradiction, suppose that $\mathrm{e} = p/q \in \mathbf{Q}$. Multiplying

$$\frac{p}{q} = \mathrm{e} = \frac{1}{0!} + \frac{1}{1!} + \cdots + \frac{1}{q!} + \frac{1}{(q+1)!} + \cdots$$

by $q!$, we obtain equality

$$p \cdot (q-1)! = \left(\frac{q!}{0!} + \frac{q!}{1!} + \cdots + \frac{q!}{q!}\right) + \sum_{n \geq q+1} \frac{q!}{n!} = a + b$$

where $a \in \mathbf{N}$ and, by simple estimates using geometric series, $0 < b < 1/q \leq 1$. Thus $a + b \notin \mathbf{N}$, which contradicts the fact that the left side $p \cdot (q-1)!$ is a positive integer.

The transcendence of e was proved first by Charles Hermite (1822–1901). The ingenious simple proof presented here belongs to David Hilbert (1862–1943). While the irrationality proof uses infinite series for e, Hilbert's proof of transcendence rests on the key property of exponential function: $(\mathrm{e}^x)' = \mathrm{e}^x$.

Theorem (Hermite, 1873). *The number* $\mathrm{e} = 2.71828\ldots$ *is transcendental.*

Proof (Hilbert, 1893). Suppose for the contradiction that Euler's number is algebraic:

$$a_n \mathrm{e}^n + \cdots + a_1 \mathrm{e} + a_0 = 0$$

for $n \in \mathbf{N}$ and some integers a_i, not all of them zero. Dividing by a power of e, we may assume that $a_0 \neq 0$. Multiplying this equation by the number

$$\int_0^\infty x^r((x-1)(x-2)\ldots(x-n))^{r+1}\mathrm{e}^{-x}\,dx$$

that depends on the parameter $r \in \mathbf{N}$ (which we choose appropriately later), we get

$$a_n \mathrm{e}^n \int_0^\infty + a_{n-1}\mathrm{e}^{n-1} \int_0^\infty + \ \cdots \ + a_1 \mathrm{e} \int_0^\infty + a_0 \int_0^\infty = 0.$$

Splitting the interval of integration $[0,\infty)$ in two, $[0,i]$ and $[i,\infty)$, we rearrange the last equation as

$$P_1(r) + P_2(r) = \left(\sum_{i=0}^{n} a_i \mathrm{e}^i \int_0^i\right) + \left(\sum_{i=0}^{n} a_i \mathrm{e}^i \int_i^\infty\right) = 0.$$

We shall prove that

$$|P_1(r)| < c^r \quad \text{for all } r \in \mathbf{N}$$

with a constant $c > 1$ not depending on r and that

$$|P_2(r)| \geq r! \quad \text{for infinitely many } r \in \mathbf{N}.$$

Then $P_1(r) + P_2(r) = 0$ cannot hold for every $r \in \mathbf{N}$ because $|P_1(r)|/r! \to 0$ as $r \to \infty$ but $|P_2(r)|/r! \geq 1$ for infinitely many r, and we obtain a contradiction.

We bound $P_1(r)$. On the interval $[0,n]$,

$$|x^r((x-1)(x-2)\dots(x-n))^{r+1}| \leq n^r(n^n)^{r+1} \quad \text{and} \quad |\mathrm{e}^{-x}| \leq 1.$$

Therefore, for $i = 0, 1, \dots, n$,

$$\left|\int_0^i x^r((x-1)(x-2)\dots(x-n))^{r+1}\mathrm{e}^{-x}\, dx\right| \leq i n^r(n^n)^{r+1} \leq (n^{n+1})^{r+1}$$

and

$$\begin{aligned} |P_1(r)| &= \left|\sum_{i=0}^{n} a_i \mathrm{e}^i \int_0^i\right| \leq |a_0| + |a_1|\mathrm{e}\left|\int_0^1\right| + \dots + |a_n|\mathrm{e}^n\left|\int_0^n\right| \\ &\leq (|a_0| + |a_1|\mathrm{e} + \dots + |a_n|\mathrm{e}^n)(n^{n+1})^{r+1} \end{aligned}$$

which is bound of the type c^r.

To bound $P_2(r)$, we evaluate first the integral $\int_0^\infty x^k \mathrm{e}^{-x}\, dx$ for $k \in \mathbf{N}_0$. Integration by parts gives

$$\begin{aligned} \int_0^\infty x^k \mathrm{e}^{-x}\, dx &= [-x^k\mathrm{e}^{-x}]_0^\infty + k\int_0^\infty x^{k-1}\mathrm{e}^{-x}\, dx \\ &= k\int_0^\infty x^{k-1}\mathrm{e}^{-x}\, dx = \dots = k!\int_0^\infty \mathrm{e}^{-x}\, dx \\ &= k!. \end{aligned}$$

(Integral $\int_0^\infty x^{s-1}\mathrm{e}^{-x}\,dx$ defines the gamma function $\Gamma(s)$.) More generally, if $p(x) = b_n x^n + \cdots + b_1 x + b_0$ is a polynomial, then

$$\int_0^\infty p(x)\mathrm{e}^{-x}\,dx = \sum_{k=0}^{n} b_k k!.$$

In particular, if $p(x)$ is an integral polynomial, then we have the congruence

$$\int_0^\infty x^k p(x)\mathrm{e}^{-x}\,dx \equiv b_0 k! \pmod{(k+1)!}.$$

Substituting $y = x - i$ we get

$$\begin{aligned}\mathrm{e}^i \int_i^\infty &= \int_i^\infty x^r((x-1)(x-2)\dots(x-n))^{r+1}\mathrm{e}^{-(x-i)}\,dx \\ &= \int_0^\infty (y+i)^r((y+i-1)(y+i-2)\dots(y+i-n))^{r+1}\mathrm{e}^{-y}\,dy.\end{aligned}$$

For $i = 0$ the polynomial in the integrand is $(-1)^{n(r+1)}(n!)^{r+1}y^r + ay^{r+1} + \cdots$ and for $1 \le i \le n$ the smallest power of y with nonzero coefficient in it is y^{r+1}. By the above calculations and the congruence, $P_2(r) \in \mathbf{Z}$ and

$$P_2(r) = \sum_{i=0}^{n} a_i \mathrm{e}^i \int_i^\infty \equiv a_0(-1)^{n(r+1)}(n!)^{r+1}r! \pmod{(r+1)!}.$$

Thus $P_2(r)$ is an integral multiple of $r!$ and, moreover, we claim that if $r+1$ is coprime with $a_0 \cdot n!$ then $P_2(r) \ne 0$. Indeed, if $P_2(r) = 0$ then by dividing the last congruence by $r!$ we get

$$0 \equiv \pm a_0(n!)^{r+1} \pmod{r+1},$$

which is impossible when $r+1 \ge 2$ and $r+1$ is coprime with $a_0 \cdot n!$. It is easy to find infinitely many $r \in \mathbf{N}$ for which this is true (recall that $a_0 \ne 0$). We may take, for example, all numbers $r = p - 1$ where p runs through the prime numbers not dividing $a_0 \cdot n!$. Hence $P_2(r)$ is a nonzero integral multiple of $r!$ for infinitely many r and thus $|P_2(r)| \ge r!$ for infinitely many r. □

Using similar arguments, Hilbert gave a simple proof for the transcendence of π as well (it was proved first by Ferdinand von Lindemann (1852-1939) in 1882).

1.5 Remarks

Much of the material is taken from Schmidt's excellent monograph [18]. The reader will find in it a proof of Roth's theorem and discussion of further refinements of Hurwitz's results. For the history of Farey fractions see Bruckheimer and Arcavi [6].

Chapter 2

Diophantine equations

We begin with a brief description of three great achievements of diophantine analysis in the 20th century: Hilbert's Tenth Problem, Fermat's Last Theorem and Catalan's conjecture. In section 2.2 we develop the theory of diophantine equation $x^2 - dy^2 = 1$ called Pell equation. Then we describe Thue equation, a large class of bivariate diophantine equations with only finitely many solutions (this fact we will not prove in entirety). In section 2.3 we discuss special cases of Fermat's Last Theorem for exponents 2 and 4 and formulate the analogue of FLT for polynomials. We present the surprising elementary and short proof of this variant of FLT based on Stothers-Mason theorem. The ABC conjecture is mentioned.

2.1 Three famous exproblems

Typical problem in diophantine equations is to decide, for a given integral polynomial $P(x_1, \ldots, x_n)$, if there are numbers $a_1, \ldots, a_n \in \mathbf{Z}$ such that $P(a_1, \ldots, a_n) = 0$, that is, if $P = 0$ has solution in integers. Besides this qualitative problem, one is also interested in the number of solutions, their size and properties. Many other problems arise when coefficients and/or solutions are taken from $\mathbf{Q}$ or other rings of numbers. Systems of equations and non-polynomial equations are investigated as well.

Note that a procedure deciding solvability of a single polynomial equation can be used to decide solvability of any system of polynomial equations:

System of equations

$$P_1 = 0 \,\&\, P_2 = 0 \,\&\, \ldots \,\&\, P_r = 0$$

($P_i \in \mathbf{Z}[x_1, \ldots, x_n]$) is solvable in $\mathbf{Z}$ if and only if the single equation

$$P_1^2 + P_2^2 + \cdots + P_r^2 = 0$$

is solvable in $\mathbf{Z}$. It is even possible, and we leave this as an exercise for the reader, to construct an algorithm $\mathcal{A}$ such that for every input polynomial $P \in \mathbf{Z}[x_1, \ldots, x_n]$ the output $Q := \mathcal{A}(P)$ is a polynomial $Q \in \mathbf{Z}[x_1, \ldots, x_r]$ with the properties that (i) $P = 0$ has solution in $\mathbf{Z}$ if and only if $Q = 0$ has solution in $\mathbf{Z}$ and (ii) $\deg(Q) \leq 4$.

Hilbert's Tenth Problem. A natural question is if there is an algorithm that would decide, for any input polynomial $P \in \mathbf{Z}[x_1, \ldots, x_n]$, if $P = 0$ has a solution in $\mathbf{Z}$. To provide such an algorithm was required in the tenth problem from the famous list of 23 mathematical problems presented by David Hilbert on the International Congress of Mathematicians in Paris in 1900. In 1970 Yuri Vladimirovich Matiyasevich (1946) [1] proved that no such algorithm exists.

Theorem (Matiyasevich, 1970). *The problem of solvability of integral polynomial equations in integers is algorithmically undecidable.*

The most famous problem in diophantine equations, and perhaps in all mathematics, was *Fermat's Last Theorem* (FLT). Fermat claimed that he could prove that equation $x^n + y^n = z^n$ has for $n \geq 3$ no solution in positive integers x, y, z. (For $n = 2$ there are infinitely many solutions, for example $3, 4, 5$ or $5, 12, 13$.) FLT was proved by Andrew Wiles (1953), with the help of Richard Taylor.

Theorem (Wiles, 1995). *For $n \geq 3$ the diophantine equation $x^n + y^n = z^n$ has no solution $x, y, z \in \mathbf{N}$.*

Considerably less known in the public than FLT was *Catalan's Conjecture*. In 1844, Eugène Catalan (1814–1894) conjectured that the only

[1] Jurij Vladimirovič Matijasevič, Юрий Владимирович Матиясевич.

solution of $x^u - y^v = 1$ in integers bigger than 1 is $3^2 - 2^3 = 1$. That is, if we mark in the sequence of natural numbers all pure powers a^b, $a, b \geq 2$,

$$1, 2, 3, \mathbf{4}, 5, 6, 7, \mathbf{8}, \mathbf{9}, 10, 11, 12, 13, 14, 15, \mathbf{16}, 17, 18, 19, 20, 21, 22, 23, 24, \mathbf{25}, 26, \mathbf{27}, 28, 29, 30, 31, \mathbf{32}, 33, 34, 35, \mathbf{36}, 37, 38, \ldots,$$

then $8, 9$ is the only pair of consecutive marked numbers. Many partial results towards the conjecture were found. Recently it was fully resolved by Preda Mihăilescu (1955).

Theorem (Mihăilescu, 2004). *If $x^u - y^v = 1$ for $x, y, u, v \in \mathbf{N}$ with $x, y, u, v > 1$ then $x = 3, y = 2, u = 2, v = 3$.*

What about the equation $x^u - y^v = 2$? Is $25, 27$ the only pair of marked numbers differing by 2? At the time of writing, this is an open question. It is not even known if $x^u - y^v = 2$ has only finitely many solutions in natural numbers.

2.2 Pell equation

The first topic we shall discuss is *Pell equation*. This is diophantine equation

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

where $x, y \in \mathbf{Z}$ are unknowns and $d \in \mathbf{N}$ is a fixed parameter that is not a square. We have always solution $(x, y) = (\pm 1, 0)$ which is called *trivial solution*. (If $d = e^2 \in \mathbf{N}$ is a square, factorization $x^2 - dy^2 = (x - ey)(x + ey)$ shows that there is only trivial solution. The same holds if $d \in \mathbf{Z}$ with $d < 0$. If $d = 0$, all solutions are $(\pm 1, z)$, $z \in \mathbf{Z}$.) For small d one can find nontrivial solutions by trial and error: $(3, 2)$ for $x^2 - 2y^2 = 1$, $(2, 1)$ for $x^2 - 3y^2 = 1$, $(9, 4)$ for $x^2 - 5y^2 = 1$, $(5, 2)$ for $x^2 - 6y^2 = 1$ and so on.

We show that if there is a nontrivial solution, then there must be infinitely many of them. We claim that for any two solutions $(a, b), (e, f) \in \mathbf{Z}^2$ of $x^2 - dy^2 = 1$, the pair $(g, h) \in \mathbf{Z}^2$ defined by

$$g + h\sqrt{d} = (a + b\sqrt{d})(e + f\sqrt{d})$$

is a solution as well. To see this, note that then also $g - h\sqrt{d} = (a - b\sqrt{d})(e - f\sqrt{d})$ and therefore

$$g^2 - dh^2 = (g + h\sqrt{d})(g - h\sqrt{d})$$

$$
\begin{aligned}
&= (a+b\sqrt{d})(e+f\sqrt{d})(a-b\sqrt{d})(e-f\sqrt{d}) \\
&= (a^2-db^2)(e^2-df^2) \\
&= 1\cdot 1 = 1.
\end{aligned}
$$

Thus if we have a nontrivial solution $(a,b) \in \mathbf{N}^2$ then for $k = 1, 2, \ldots$

$$a_k + b_k\sqrt{d} = (a+b\sqrt{d})^k$$

give infinitely many solutions $(a_k, b_k) \in \mathbf{N}^2$. For example, the solution $(a,b) = (3,2)$ of $x^2 - 2y^2 = 1$ generates solutions $(a_2, b_2) = (17, 12)$, $(a_3, b_3) = (99, 70)$ and so on. Later we shall prove that the smallest nontrivial natural solution (a,b) generates all natural solutions. All nontrivial integral solutions are obtained simply by adding signs to the natural solutions.

Theorem (Lagrange, 1770). *Every Pell equation $x^2 - dy^2 = 1$ has a nontrivial solution (and hence infinitely many solutions).*

Proof. Because d is not a square, $\sqrt{d}$ is irrational and by part 2 of Dirichlet's theorem in chapter 1 there are infinitely many distinct fractions p/q such that

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

These fractions satisfy

$$
\begin{aligned}
|p^2 - dq^2| &= q|\sqrt{d} - p/q| \cdot |p + q\sqrt{d}| < \frac{|p+q\sqrt{d}|}{q} \le p/q + \sqrt{d} \\
&\le 2\sqrt{d} + 1.
\end{aligned}
$$

Thus (by the pigeonhole principle used for infinitely many pigeons and finitely many holes) there is a $c \in \mathbf{Z}$ such that $p^2 - dq^2 = c$ for infinitely many $p/q \in \mathbf{Q}$. Irrationality of $\sqrt{d}$ implies that $c \neq 0$. There are only finitely many, c^2, possibilities for the residues of the pairs p, q modulo $|c|$. Thus we can select two distinct fractions $p_1/q_1, p_2/q_2$ such that $p_1^2 - dq_1^2 = p_2^2 - dq_2^2 = c$ and $p_1 \equiv p_2$, $q_1 \equiv q_2$ modulo $|c|$ (in fact, there are infinitely many such fractions). Consider the numbers a, b defined by

$$
\begin{aligned}
a + b\sqrt{d} &= \frac{p_1 + q_1\sqrt{d}}{p_2 + q_2\sqrt{d}} = \frac{(p_1 + q_1\sqrt{d})(p_2 - q_2\sqrt{d})}{(p_2 + q_2\sqrt{d})(p_2 - q_2\sqrt{d})} \\
&= \frac{p_1p_2 - dq_1q_2}{c} + \frac{p_2q_1 - p_1q_2}{c}\sqrt{d}.
\end{aligned}
$$

We claim that (a, b) is a nontrivial solution of $x^2 - dy^2 = 1$.

The numerators are integral multiples of c because, using the congruences $p_1 \equiv p_2$ and $q_1 \equiv q_2$, modulo $|c|$ we have $p_1p_2 - dq_1q_2 \equiv p_1^2 - dq_1^2 = c \equiv 0$ and $p_2q_1 - p_1q_2 \equiv p_1q_2 - p_1q_2 = 0$. Thus $a, b \in \mathbf{Z}$. We cannot have $b = 0$ because $p_1/q_1 \neq p_2/q_2$. Thus $b \neq 0$. Finally,

$$\begin{aligned} a^2 - db^2 &= (a + b\sqrt{d})(a - b\sqrt{d}) = \frac{p_1 + q_1\sqrt{d}}{p_2 + q_2\sqrt{d}} \cdot \frac{p_1 - q_1\sqrt{d}}{p_2 - q_2\sqrt{d}} \\ &= \frac{p_1^2 - dq_1^2}{p_2^2 - dq_2^2} = \frac{c}{c} = 1. \end{aligned}$$

$\square$

It is convenient to view solutions of $x^2 - dy^2 = 1$ as real numbers and collect them in the set

$$R = \{a + b\sqrt{d} :\ a, b \in \mathbf{Z}, a^2 - db^2 = 1\}.$$

For $\alpha = a + b\sqrt{d} \in R$ we denote $\overline{\alpha} = a - b\sqrt{d}$ and have $\alpha\overline{\alpha} = 1$. Thus R is closed to division. We have seen above that R is closed to multiplication. Hence $(R, \cdot)$ is a multiplicative abelian group with the neutral element $1 = 1 + 0\sqrt{d}$. By Lagrange's theorem there exists $\alpha \in R$ with $\alpha > 1$ and therefore R is infinite (α^k are distinct for $k = 1, 2, \ldots$ and lie in R).

For $\alpha \in R$ the numbers α and $\overline{\alpha}$ have the same sign and are separated (if they are different from ± 1) either by -1 or by 1. The four possibilities for the signs of a, b in $\alpha = a + b\sqrt{d}$, $\alpha \neq \pm 1$, therefore determine whether α lies in $R \cap (-\infty, -1)$ (signs are $-, -$), in $R \cap (-1, 0)$ (signs are $-, +$), in $R \cap (0, 1)$ (signs are $+, -$) or in $R \cap (1, \infty)$ (signs are $+, +$). Thus the subgroup of positive solutions

$$U = \{\alpha \in R :\ \alpha > 0\}$$

is formed by the $a + b\sqrt{d} \in R$ with $a \in \mathbf{N}$, and the natural solutions $\alpha = a + b\sqrt{d} \in R$ with $a, b \in \mathbf{N}$ are exactly the solutions $R \cap (1, \infty)$.

If $\alpha = a_1 + b_1\sqrt{d}, \beta = a_2 + b_2\sqrt{d}$ are two natural solutions then $a_1 < a_2 \iff b_1 < b_2 \iff \alpha < \beta$. Now it is clear that there exists the smallest natural solution

$$\varepsilon = \varepsilon(d) = \min\{\alpha \in U :\ \ \alpha > 1\}.$$

We claim that $U = \{\varepsilon^k :\ k \in \mathbf{Z}\}$, that is, $(U, \cdot)$ is a cyclic group generated by ε. Let $\alpha \in U$. If $\alpha > 1$, there is a $k \in \mathbf{N}$ such that $\varepsilon^k \leq \alpha < \varepsilon^{k+1}$.

If $\varepsilon^k < \alpha$, dividing by ε^k we get $1 < \alpha\varepsilon^{-k} < \varepsilon$ and $\alpha\varepsilon^{-k} \in U \cap (1, \infty)$, which contradicts the minimality of ε. Thus $\alpha = \varepsilon^k$. If $0 < \alpha \leq 1$, we have $\alpha^{-1} \geq 1$ and $\alpha = \varepsilon^{-k}$ for some $k \in \mathbf{N}_0$. We summarize our findings.

Theorem. *The set of positive solutions of Pell equation $x^2 - dy^2 = 1$ is an infinite cyclic group $(U, \cdot)$ that is generated by the smallest natural solution ε. The mapping $\varepsilon^k \mapsto k$ is a group isomorphism between $(U, \cdot)$ and $(\mathbf{Z}, +)$. The group of all solutions $(R, \cdot)$ is isomorphic to $(\mathbf{Z}, +) \times (\mathbf{Z}_2, +)$.* □

The *generalized Pell equation* is the diophantine equation

$$x^2 - dy^2 = m$$

where $x, y \in \mathbf{Z}$ are unknowns and $d \in \mathbf{N}, m \in \mathbf{Z}$ are parameters and d is not a square.

Theorem. *If the generalized Pell equation $x^2 - dy^2 = m$ has an integral solution, then it has infinitely many integral solutions.*

Proof. Suppose $a, b \in \mathbf{Z}$ is a solution of $x^2 - dy^2 = m$ and $e, f \in \mathbf{Z}$ is a solution of $x^2 - dy^2 = 1$. Then

$$g + h\sqrt{d} = (a + b\sqrt{d})(e + f\sqrt{d})$$

is a solution of $x^2 - dy^2 = m$ as well because

$$\begin{aligned} g^2 - dh^2 &= (g + h\sqrt{d})(g - h\sqrt{d}) \\ &= (a + b\sqrt{d})(e + f\sqrt{d})(a - b\sqrt{d})(e - f\sqrt{d}) \\ &= (a^2 - db^2)(e^2 - df^2) \\ &= m \cdot 1 = m. \end{aligned}$$

Multiplying one solution of $x^2 - dy^2 = m$ by infinitely many solutions of $x^2 - dy^2 = 1$ we get infinitely many solutions of $x^2 - dy^2 = m$. □

2.3 Thue equation

We proceed to other classical family of diophantine equations. *Thue equation* is a diophantine equation

$$F(x, y) = m$$

where x, y are unknowns, $m \in \mathbf{Z}$ is a parameter, $m \neq 0$, and $F \in \mathbf{Z}[x, y]$ is a nonzero homogeneous integral polynomial that is irreducible over $\mathbf{Z}[x, y]$ and has degree $n \geq 3$.

We will not be able to give a complete proof of the following breakthrough result of Axel Thue. We only show how it follows from Thue's theorem on algebraic numbers (chapter 1).

Theorem (Thue, 1909). *Every Thue equation $F(x, y) = m$ has only finitely many integral solutions $x, y \in \mathbf{Z}$.*

We have

$$F(x, y) = a_n x^n + a_{n-1}x^{n-1}y + \cdots + a_1 xy^{n-1} + a_0 y^n = y^n F(x/y)$$

where

$$F(z) = a_n z^n + a_{n-1}z^{n-1} + \cdots + a_0.$$

Since $a_0, a_n \neq 0$ ($F(x, y)$ is irreducible), $\deg(F(z)) = \deg(F(x, y)) = n \geq 3$. The irreducibility of $F(x, y)$ is equivalent with the irreducibility of $F(z)$. The irreducibility of $F(z)$ over $\mathbf{Z}[x]$ is equivalent with the irreducibility over $\mathbf{Q}[x]$ (Gauss' lemma). The irreducibility over $\mathbf{Q}[x]$ implies that in the factorization

$$F(z) = a_n \prod_{i=1}^{n}(z - \alpha_i)$$

all algebraic numbers $\alpha_i \in \mathbf{C}$ are distinct and have degree $n \geq 3$.

Proof. Suppose that $F(p, q) = m$ for infinitely many pairs $(p, q) \in \mathbf{Z}^2$. For each q only at most n numbers $p \in \mathbf{Z}$ satisfy $F(p, q) = m$. Thus there are infinitely many pairs $(p, q) \in \mathbf{Z}^2$ satisfying $F(p, q) = m$ in which the q's are mutually distinct and nonzero. Using the factorization of $F(z)$ and dividing by q^n, we get

$$\prod_{i=1}^{n}\left(\frac{p}{q} - \alpha_i\right) = \frac{m}{a_n q^n}.$$

Changing signs of the p's we can assume that $q \in \mathbf{N}$. We let $q \to \infty$.

Let $v = \min|\alpha_i - \alpha_j| > 0$ for $i \neq j$ (here we use that all α_i are distinct). For big enough q the right side in the last equation is in absolute value smaller than $(v/2)^n$, which forces

$$\left|\alpha_i - \frac{p}{q}\right| < v/2$$

for some i. By the pigeonhole principle, there is an i_0 for which this happens infinitely many times; let it be $i_0 = 1$. By the triangle inequality, for $j \geq 2$ we have

$$\left|\alpha_j - \frac{p}{q}\right| \geq |\alpha_j - \alpha_1| - \left|\alpha_1 - \frac{p}{q}\right| > v - \frac{v}{2} = v/2.$$

Thus

$$\left|\alpha_1 - \frac{p}{q}\right| = \prod_{j=2}^{n} \left|\alpha_j - \frac{p}{q}\right|^{-1} \left|\frac{m}{a_n q^n}\right| < \frac{c}{q^n}$$

where $c = |m/a_n| \cdot (2/v)^{n-1} > 0$ is a constant. Since α_1 is an algebraic number with degree $n \geq 3$ and this inequality holds for infinitely many fractions p/q, for big enough q we get a contradiction with Thue's inequality stated in chapter 1. $\square$

It is easy to see that $F(x, y) = x^3 - 2y^3$ is irreducible and therefore, by the last theorem, every equation $x^3 - 2y^3 = m$, $m \in \mathbf{Z}$, has only finitely many integral solutions. This contrasts with Pell equation $x^2 - 2y^2 = 1$ that has infinitely many solutions.

2.4 FLT for $n = 2, 4$ and for polynomials

Our last topic in the second chapter is FLT for exponents $n = 2$ and $n = 4$ and FLT for polynomials. For exponent 2 we have the diophantine equation

$$x^2 + y^2 = z^2$$

and consider its nontrivial solutions $x, y, z \in \mathbf{Z}$, $xyz \neq 0$. Common factor can be canceled out and signs can be changed so that x, y, z are pairwise coprime natural numbers. Numbers x, y cannot be both even and they cannot be both odd either (consider squares modulo 4). We may assume that x is even and y is odd. We call triples of numbers with these properties—$(a, b, c) \in \mathbf{N}^3$ such that $a^2 + b^2 = c^2$, a, b, c are coprime, a is even and b is odd—*primitive Pythagorean triples*. Every nontrivial integral solution of $x^2 + y^2 = z^2$ is obtained from a primitive Pythagorean triple by a change of signs, multiplication by a common integral factor, and switching x and y.

Theorem. *A triple $(x, y, z) \in \mathbf{N}^3$ is a primitive Pythagorean triple if and only if $x = 2ab$, $y = a^2 - b^2$, and $z = a^2 + b^2$ for some integers $a > b \geq 1$ that* *are coprime and have different parity. In particular, there exist infinitely many primitive Pythagorean triples.*

Proof. It is easy to verify that the numbers given by the formulae form a primitive Pythagorean triple. Note that $(2ab)^2 + (a^2 - b^2)^2 = (a^2 + b^2)^2$ is in fact a polynomial identity. On the other hand, if $(x, y, z) \in \mathbf{N}^3$ is a primitive Pythagorean triple, we rewrite $x^2 + y^2 = z^2$ as

$$\left(\frac{x}{2}\right)^2 = \frac{z-y}{2} \cdot \frac{z+y}{2}$$

and deduce that both factors on the right, which are natural numbers (z, y are odd), are squares. Indeed, they are coprime (because their sum is z and their difference y and $(z, y) = 1$) and their product is a square. Thus $(z - y)/2 = b^2$ and $(z + y)/2 = a^2$ for some natural numbers $a > b$. Adding and subtracting we get $z = a^2 + b^2$ and $y = a^2 - b^2$. Also, $x = 2ab$ and it follows that $(a, b) = 1$ and a, b have different parity. $\square$

Theorem (de Fermat, 17-th century). *The diophantine equation*

$$x^4 + y^4 = z^2$$

has no solution $x, y, z \in \mathbf{N}$.

Proof. Suppose $(x, y, z) \in \mathbf{N}^3$ is a solution. We may assume that x, y, z are coprime. (The common factor can be divided out so that a solution with coprime components is obtained.) Again, x, y have different parity and we assume that x is odd and y is even. We rewrite the equation as

$$y^4 = (z - x^2)(z + x^2).$$

Since z, x are odd, $(z, x) = 1$, and the sum and the difference of the factors is $2z$ and $2x^2$, we see that $(z - x^2, z + x^2) = 2$. This implies that

$$z - x^2 = 2a^4 \quad \text{and} \quad z + x^2 = 8b^4,$$

or the right sides are switched, and a, b are coprime and a is odd. Subtracting the equations we get $x^2 = 4b^4 - a^4$, which is impossible modulo 4. Thus the right sides must be switched:

$$z - x^2 = 8b^4 \quad \text{and} \quad z + x^2 = 2a^4$$

and $x^2 = a^4 - 4b^4$, $z = a^4 + 4b^4$. We rewrite the former equation as

$$4b^4 = (a^2 - x)(a^2 + x).$$

It follows again that $(a^2 - x, a^2 + x) = 2$. Now we have only one possibility that each factor is twice a biquadrate: $a^2 - x = 2c^4$, $a^2 + x = 2d^4$ with $c, d \in \mathbf{N}$. Adding both equations we get $a^2 = c^4 + d^4$. Starting from the solution $(x, y, z) \in \mathbf{N}^3$, we have constructed another solution $(c, d, a) \in \mathbf{N}^3$ of the same equation. But $a < z$ (because $z = a^4 + 4b^4$). Repeating the argument, we could obtain infinitely many natural solutions whose third components would form an infinite strictly descending sequence. This is in the set $\mathbf{N}$ impossible and we arrive at a contradiction. Thus there is no solution in natural numbers. □

The concluding argument, which shows nonexistence of a solution by constructing an infinite strictly descending sequence of natural numbers, is called the *infinite descend argument* and was invented by Fermat.

What about FLT for polynomials? It holds for them and it can be proved by an ingenious argument that is much shorter than the 100+ pages argument of Wiles for numbers. To state the key result, we need a definition. If $p(t)$ is a complex polynomial then $r(p)$ denotes the number of its *distinct* roots. For example, $r(t^4 - 2t^2 + 1) = 2$. Obviously, $r(p) \leq \deg(p)$ and $r(p^n) = r(p)$ for every $n \in \mathbf{N}$.

Theorem (Stothers, 1981; Mason, 1984). *If*

$$a(t) + b(t) = c(t)$$

for coprime polynomials $a, b, c \in \mathbf{C}[t]$ which are not all constant, then

$$\max(\deg(a), \deg(b), \deg(c)) \leq r(abc) - 1.$$

First we show how the polynomial FLT follows from this.

Theorem. *If*

$$x(t)^n + y(t)^n = z(t)^n$$

for $n \in \mathbf{N}$ and coprime polynomials $x, y, z \in \mathbf{C}[t]$ which are not all constant, then $n \leq 2$.

Proof. Using the Stothers–Mason theorem and simple properties of degrees of polynomials, we get

$$\begin{aligned} n \deg(x) &= \deg(x^n) \leq r(x^n y^n z^n) - 1 = r(xyz) - 1 \leq \deg(xyz) - 1 \\ &= \deg(x) + \deg(y) + \deg(z) - 1 \end{aligned}$$

and the same upper bound on $n \deg(y)$ and $n \deg(z)$. Adding the three bounds, we obtain

$$n(\deg(x) + \deg(y) + \deg(z)) \leq 3(\deg(x) + \deg(y) + \deg(z)) - 3,$$

which implies that $n \leq 2$. □

For $n = 2$, many solutions can be obtained from the polynomial identity $(2ab)^2 + (a^2 - b^2)^2 = (a^2 + b^2)^2$, for example $(2t)^2 + (t^2 - 1)^2 = (t^2 + 1)^2$.

Proof of the Stothers-Mason theorem. Dividing $a + b = c$ by c ($c \neq 0$ by coprimality), setting $f = a/c$, $g = b/c$ and differentiating, we get the equations

$$f + g = 1 \quad \text{and} \quad f' + g' = f \cdot \frac{f'}{f} + g \cdot \frac{g'}{g} = 0.$$

The last equation can be rearranged as

$$-\frac{f'/f}{g'/g} = \frac{g}{f} = \frac{b}{a}.$$

Splitting a, b, c in the linear factors, we write

$$f = \frac{a}{c} = \frac{\alpha \prod (t - \alpha_i)^{m_i}}{\gamma \prod (t - \gamma_i)^{o_i}} \quad \text{and} \quad g = \frac{b}{c} = \frac{\beta \prod (t - \beta_i)^{n_i}}{\gamma \prod (t - \gamma_i)^{o_i}}.$$

Taking the logarithmic derivatives $f'/f = (\log f)'$, $g'/g = (\log g)'$ of these factorizations and substituting them in the above equation we get

$$\frac{b}{a} = -\frac{\sum m_i/(t - \alpha_i) - \sum o_i/(t - \gamma_i)}{\sum n_i/(t - \beta_i) - \sum o_i/(t - \gamma_i)}.$$

We multiply the denominator and the numerator of the right side by

$$N = \prod (t - \alpha_i) \cdot \prod (t - \beta_i) \cdot \prod (t - \gamma_i)$$

and get

$$\frac{b}{a} = -\frac{N\left(\sum m_i/(t-\alpha_i) - \sum o_i/(t-\gamma_i)\right)}{N\left(\sum n_i/(t-\beta_i) - \sum o_i/(t-\gamma_i)\right)} = \frac{Q}{P}$$

where P and Q are polynomials which have degrees at most $\deg(N) - 1 = r(abc) - 1$. But a, b are coprime polynomials and therefore $\deg(a) \leq \deg(P) \leq r(abc) - 1$ and $\deg(b) \leq \deg(Q) \leq r(abc) - 1$. Since $a + b = c$, we have also $\deg(c) \leq \max(\deg(a), \deg(b)) \leq r(abc) - 1$. □

An analogous statement can be made for integers but it is so far unproved; it is the famous *ABC conjecture.*

The abc conjecture (Maser and Oysterlé, 1985). *Let* $\mathrm{rad}(m)$ *denote the product of all prime divisors of* $m \in \mathbf{Z}$. *For every* $\varepsilon > 0$ *there is a constant* $K = K(\varepsilon) > 0$ *such that if*

$$a + b = c$$

for coprime integers a, b, c, *then*

$$\max(|a|, |b|, |c|) \leq K \mathrm{rad}(abc)^{1+\varepsilon}.$$

Corollary (asymptotic FLT). *If the abc conjecture is true then there is an* $n_0 \in \mathbf{N}$ *such that for* $n \geq n_0$ *the equation*

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

has no solution $x, y, z \in \mathbf{N}$.

Proof. Let $x^n + y^n = z^n$ for some $x, y, z \in \mathbf{N}$. Since $\max(x^n, y^n, z^n) = z^n$ and $\mathrm{rad}(x^n y^n z^n) = \mathrm{rad}(xyz) \leq xyz \leq z^3$, using the abc conjecture with $\varepsilon = 1$ we get

$$z^n \leq K \mathrm{rad}(x^n y^n z^n)^2 \leq K z^6$$

with a constant K which can be taken bigger than 1. Taking logarithms we have

$$n \leq \frac{\log K}{\log z} + 6 \leq \frac{\log K}{\log 2} + 6$$

because $\log K > 0$ and $z \geq 2$. Thus we may set $n_0 = \lceil \log K / \log 2 \rceil + 7$. □

Of course, Wiles proved the FLT unconditionally for all $n \geq 3$ but the simplicity of the derivation shows the strength of the abc conjecture. Many other consequences were derived from it.

2.5 Remarks

Thorough presentation of Matiyasevich's achievement was given by Davis [8], see also Jones and Matijasevič [13] and Matijasevič [15]. The proof of Fermat's Last Theorem due to Wiles was published in [19] and [21]. For the solution of Catalan's conjecture see Mihăilescu [16] and expositions [4] and [5] by Bilu. In the part on Pell equation we follow Hlawka, Schoißengeier and Taschner [10] and in that on FLT for polynomials Lang [14]. For more information on the abc conjecture see [22].

Chapter 3

Geometry of numbers

In this chapter we present some arguments based on geometry and show their applications in number theory. We prove Minkowski's theorem, which says that centrally symmetric convex body with large volume must contain many lattice points, and deduce from it Lagrange's theorem asserting that every natural number is a sum of at most four squares. We present also an arithmetic proof of Lagrange's theorem. A geometric proof of the main property of Farey fractions (it says that two consecutive Farey fractions have, in an appropriate sense, smallest possible distance) is given. In conclusion we discuss the asymptotic behaviors of the number of lattice points lying in a circle and of the number of lattice points lying under a branch of hyperbola.

3.1 Lattices, Farey fractions and convex bodies

A *lattice* $\Lambda = \Lambda(B)$ in $\mathbf{R}^n$ with the *base* $B = \{v_1, v_2, \ldots, v_n\}$, where v_i are n linearly independent vectors in $\mathbf{R}^n$, is the set of all integral linear combinations of the vectors in the base:

$$\Lambda = \{\textstyle\sum_1^n a_i v_i : \ a_i \in \mathbf{Z}\}.$$

The *fundamental parallelepiped* $T = T(B)$ of $\Lambda(B)$ is the set

$$T = \{\textstyle\sum_1^n \alpha_i v_i : \ \alpha_i \in [0,1)\}.$$

Every $u \in \mathbf{R}^n$ has a unique expression as $u = \sum b_i v_i$ with $b_i \in \mathbf{R}$ and hence a unique expression as $u = \sum a_i v_i + \sum \alpha_i v_i$ with $a_i \in \mathbf{Z}$ and $\alpha_i \in [0,1)$ (write $b_i = \lfloor b_i \rfloor + \{b_i\}$). Thus the system

$$\{z + T :\ z \in \Lambda\}$$

of translated copies of T (here $z + T = \{z + t :\ t \in T\}$) is a set partition of $\mathbf{R}^n$, which means that each vector $u \in \mathbf{R}^n$ lies in exactly one set $z + T$.

Recall that the volume $\mathrm{Vol}(T)$ of $T = T(B)$ can be calculated as the absolute value of the determinant of the matrix $M(B)$ whose rows are the vectors in B:

$$\mathrm{Vol}(T) = |\det(M(B))|.$$

One lattice Λ has many bases and many fundamental parallelepipeds. We show that their volumes are all equal and define a constant $\mathrm{Vol}(\Lambda)$, the *volume of the lattice* Λ.

Proposition. *If B_1, B_2 are two bases of a lattice Λ and T_1, T_2 are the corresponding fundamental parallelepipeds then*

$$\mathrm{Vol}(T_1) = \mathrm{Vol}(T_2).$$

Proof. Let $B_1 = \{v_1, v_2, \ldots, v_n\}$, $B_2 = \{w_1, w_2, \ldots, w_n\}$ and $V = M(B_1)$ and $W = M(B_2)$ be the matrices of the bases (vectors of the bases are in the rows). We express vectors of one base as integral linear combinations of the vectors in the other base: $v_i = \sum_j a_{ij} w_j$ and $w_i = \sum_j b_{ij} v_j$ where a_{ij} and b_{ij} are integers. In matrix notation,

$$V = AW \quad \text{and} \quad W = BV$$

where $A = (a_{ij})$ and $B = (b_{ij})$. Matrices V, W are regular. We write $A = VW^{-1}, B = WV^{-1}$ and obtain that $AB = BA = E$, that is, A and B are inverses of one another. Multiplication of determinants gives

$$\det(A)\det(B) = \det(AB) = \det(E) = 1.$$

But $\det(A), \det(B) \in \mathbf{Z}$ because A, B are integral matrices. Thus $\det(A) = \det(B) = \pm 1$ and

$$\mathrm{Vol}(T_1) = |\det(V)| = |\det(A)| \cdot |\det(W)| = |\det(W)| = \mathrm{Vol}(T_2).$$

$\square$

One of the simplest and most important examples of a lattice is $\mathbf{Z}^2 = \{(a,b): \ a,b \in \mathbf{Z}\}$, the lattice of the *lattice points* in the plane $\mathbf{R}^2$. One of its bases is the canonical base $\{(1,0),(0,1)\}$ and so $\mathrm{Vol}(\mathbf{Z}^2) = 1$. Using properties of lattices we give a geometric proof of Cauchy's theorem on Farey fractions from chapter 1. We show that any two consecutive Farey fractions $a/b < c/d$ form a basis $\{(a,b),(c,d)\}$ of $\mathbf{Z}^2$.

Another proof of Farey–Cauchy theorem. We prove that if $a/b < c/d$ are two fractions from the interval $[0,1]$ which are in lowest terms, have denominators $b,d \le n$, and $a/b < e/f < c/d$ holds for no other fraction e/f with $f \le n$, then

$$bc - ad = 1.$$

Consider the plane vectors $u = (a,b)$ and $v = (c,d)$. We claim that the triangle Δ with the vertices $(0,0)$, u, and v contains no lattice points besides its vertices. To show it, we write

$$\Delta = \{w \in \mathbf{R}^2: \ w = \alpha u + \beta v, 0 \le \alpha, \beta \le 1, \alpha + \beta \le 1\}.$$

Let $(e,f) \in \Delta \cap \mathbf{Z}^2$ and $(e,f) \ne (0,0)$. Then $a/b \le e/f \le c/d$. If $e/f = a/b$ then $(e,f) = (a,b)$ because a,b are coprime. If the other equality occurs we have similarly $(e,f) = (c,d)$. If $a/b < e/f < c/d$ then we have a contradiction with the properties of a/b and c/d because $f = \alpha b + \beta d \le (\alpha+\beta)n \le n$.

We complete Δ to the parallelogram

$$P = \{\alpha u + \beta v: \ \alpha, \beta \in [0,1]\} = \Delta \cup \Delta' \ \text{ where } \ \Delta' = u + v - \Delta.$$

From $\Delta \cap \mathbf{Z}^2 = \{(0,0), u, v\}$ it follows that $\Delta' \cap \mathbf{Z}^2 = \{u, v, u+v\}$ and $P \cap \mathbf{Z}^2 = \{(0,0), u, v, u+v\}$. Hence for the fundamental parallelogram

$$T = \{\alpha u + \beta v: \ \alpha, \beta \in [0,1)\}$$

of the lattice

$$\Lambda = \Lambda(\{u,v\})$$

we have $T \cap \mathbf{Z}^2 = \{(0,0)\}$. It follows that the translates of T by the vectors $z \in \Lambda$ partition $\mathbf{R}^2$ and that $\Lambda = \mathbf{Z}^2$. Thus $\mathrm{Vol}(\Lambda) = \mathrm{Vol}(\mathbf{Z}^2) = 1$,

$$|\det(M(\{u,v\}))| = |ad - bc| = 1,$$

and $bc - ad$=1. $\square$

A *convex* set $B \subset \mathbf{R}^n$, usually called a *convex body*, contains with every two of its points also the segment connecting them. B is *centrally symmetric* if $x \in B$ implies $-x \in B$ and B is *bounded* if it lies in a ball. We shall consider only measurable sets which have defined volume. The next basic result in the geometry of numbers was obtained by Hermann Minkowski (1864–1909).

Theorem (Minkowski, 1891). *If $B \subset \mathbf{R}^n$ is a convex body that is bounded and centrally symmetric and $\Lambda \subset \mathbf{R}^n$ is a lattice satisfying $2^n\mathrm{Vol}(\Lambda) < \mathrm{Vol}(B)$, then*

$$B \cap \Lambda \neq \{(0, 0, \ldots, 0)\},$$

that is, B contains a point of the lattice different from the origin .

Proof. Let T be the fundamental parallelepiped of Λ, $B_z = T \cap (\frac{1}{2}B + z)$, and $C_z = (T - z) \cap \frac{1}{2}B$. Clearly, $\mathrm{Vol}(B_z) = \mathrm{Vol}(C_z)$ because $C_z = B_z - z$ and volume is shift-invariant. We have

$$\sum_{z\in\Lambda} \mathrm{Vol}(B_z) = \sum_{z\in\Lambda} \mathrm{Vol}(C_z) = \mathrm{Vol}(\tfrac{1}{2}B) = 2^{-n}\mathrm{Vol}(B) > \mathrm{Vol}(\Lambda) = \mathrm{Vol}(T)$$

where the second equality follows from the fact that the sets C_z partition $\frac{1}{2}B$. Both sums have only finitely many nonzero summands because $\frac{1}{2}B$ intersects only finitely many translates $T - z$, $z \in \Lambda$.

(If $T = T(\{v_1, v_2, \ldots, v_n\})$ then T certainly lies in the origin-centered ball $K(r)$ with the radius $r = |v_1| + |v_2| + \cdots + |v_n|$. Since B is bounded, $K(R) \supset \frac{1}{2}B$ for some $R > 0$. Thus every $T - z$, $z \in \Lambda$, intersecting $\frac{1}{2}B$ must lie in the ball $K(R + r)$ and there are at most $\mathrm{Vol}(K(R + r))/\mathrm{Vol}(T)$ such translates because they partition $\mathbf{R}^n$.)

Since $B_z \subset T$ for every z and the sum of the volumes $\mathrm{Vol}(B_z)$ exceeds $\mathrm{Vol}(T)$, the sets B_z, $z \in \Lambda$, cannot be pairwise disjoint. This means that for two different points $z_1, z_2 \in \Lambda$ we have $(\frac{1}{2}B + z_1) \cap (\frac{1}{2}B + z_2) \neq \emptyset$. Thus $\frac{1}{2}x_1 + z_1 = \frac{1}{2}x_2 + z_2$ for two points $x_1, x_2 \in B$. This gives a point of Λ that differs from the origin and lies in B:

$$B \ni \tfrac{1}{2}(x_1 - x_2) = z_2 - z_1 \in \Lambda.$$

(The point $\frac{1}{2}(x_1 - x_2)$ is in B because it is the center of the segment connecting the points x_1 and $-x_2$ lying in B, and it differs from the origin because $z_1 \neq z_2$.) $\square$

3.2 Four-squares theorem

We shall use Minkowski's theorem to prove a famous result belonging to diophantine equations (and to additive number theory), Lagrange's *Four Squares Theorem.*

Theorem (Lagrange, 1770). *For every number $n \in \mathbf{N}_0$ the equation*

$$n = x_1^2 + x_2^2 + x_3^2 + x_4^2$$

has a solution $x_1, x_2, x_3, x_4 \in \mathbf{Z}$.

Lemma. *For every prime p the congruence*

$$a^2 + b^2 + 1 \equiv 0 \pmod{p}$$

has a solution $a, b \in \mathbf{Z}$.

Proof. For $p = 2$ we have solution $1, 0$ and we may therefore assume that $p > 2$. For $a = 0, 1, \ldots, (p-1)/2$, the numbers a^2 are mutually incongruent modulo p ($a_1^2 \equiv a_2^2$ is equivalent with $(a_1 - a_2)(a_1 + a_2) \equiv 0$ which is equivalent with $a_1 \equiv \pm a_2$) and the same holds for the numbers $-a^2 - 1$. Together we have

$$(p+1)/2 + (p+1)/2 = p + 1 > p$$

residues modulo p and a residue must be represented in both ways: $a^2 \equiv -b^2 - 1$ for some $a, b \in \{0, 1, \ldots, (p-1)/2\}$. $\square$

Corollary. *For every squarefree number $n = p_1 p_2 \ldots p_r$ the congruence*

$$a^2 + b^2 + 1 \equiv 0 \pmod{n}$$

has a solution $a, b \in \mathbf{Z}$.

Proof. Recall the Chinese remainder theorem (CHRT): If $m_1, \ldots, m_r \in \mathbf{N}$ are pairwise coprime moduli and $M = m_1 m_2 \ldots m_r$, then for every r-tuple of integers $a_1, \ldots, a_r$ the system of congruences $x \equiv a_i \pmod{m_i}$, $1 \leq i \leq r$, has exactly one solution in the set $1, 2, \ldots, M$. By the lemma there are a_i and b_i satisfying the congruence modulo p_i. By the CHRT there are integers a and b which are modulo p_i equal a_i and b_i, respectively, and thus $a^2 + b^2 + 1 \equiv 0 \pmod{p_i}$ for every $i = 1, \ldots, r$. But this implies (p_i are distinct) that $a^2 + b^2 + 1 \equiv 0 \pmod{p_1 p_2 \ldots p_r}$. $\square$

Numbers of the form $8n+7$ are not sums of three squares because modulo 8 squares produce only residues $0, 1, 4$. It can be shown that, more generally, numbers of the form $4^r(8n+7)$ are not sums of three squares either. Gauss proved that every other number is a sum of three squares.

Geometric proof of the four squares theorem. It suffices to prove it only for squarefree numbers $n = p_1 p_2 \dots p_r$ (every $n \in \mathbf{N}$ has a unique expression as $n = s^2 m$ with squarefree m and then $m = \sum_1^4 x_i^2$ gives $n = \sum_1^4 (s x_i)^2$). Let n be a squarefree number. Using the Corollary we take $a, b \in \mathbf{N}$ such that $a^2 + b^2 + 1$ is a multiple of n. We shall work in $\mathbf{R}^4$ with the lattice

$$\Lambda = \Lambda(\{u_1, u_2, u_3, u_4\})$$

where

$$u_1 = (n, 0, 0, 0),\ u_2 = (0, n, 0, 0),\ u_3 = (a, b, 1, 0),\ u_4 = (b, -a, 0, 1).$$

It is clear that u_i are linearly independent and that $\mathrm{Vol}(\Lambda) = n^2$ because the matrix of the base is lower triangular.

The second thing we need for Minkowski's theorem is a convex body B. We set $B = K(r)$, the origin-centered four-dimensional ball with radius $r > 0$. Since

$$\mathrm{Vol}(K(1)) = \pi^2/2$$

(we take this formula for granted and shall not prove it by integration), $\mathrm{Vol}(K(r)) = \pi^2 r^4/2$. The condition of Minkowski's theorem on volumes requires

$$\frac{\pi^2 r^4}{2} > 2^4 n^2,$$

which is the same as

$$r^2 > \frac{4\sqrt{2}}{\pi} n = (1.80063\dots)n.$$

We take an r defined by the equation

$$r^2 = 1.9n.$$

The condition on volumes is then satisfied and other conditions on $B = K(r)$ (convexity and central symmetry) are satisfied as well. By Minkowski's theorem there exists a $z = \sum_1^4 a_i u_i \in \Lambda$, with not all $a_i \in \mathbf{Z}$ equal to zero, that lies in $K(r)$. In terms of coordinates,

$$0 < |z|^2 = (a_1 n + a_3 a + a_4 b)^2 + (a_2 n + a_3 b - a_4 a)^2 + a_3^2 + a_4^2 \le r^2 < 2n.$$

Thus the integer $|z|^2$ is a sum of four squares and lies in the interval $[1, 2n - 1]$. But

$$|z|^2 = a_3^2(a^2 + b^2 + 1) + a_4^2(a^2 + b^2 + 1) + 2a_3 a_4 ab - 2a_3 a_4 ab + n(\cdots)$$

shows that $|z|^2$ is a multiple of n (because of the selection of a and b). The only possibility is $|z|^2 = n$ and n is a sum of four squares. □

For the second proof of Lagrange's theorem we need a remarkable identity due to Euler:

Lemma (Euler's four squares identity).

$$\begin{aligned}
&(x_1^2 + x_2^2 + x_3^2 + x_4^2)(y_1^2 + y_2^2 + y_3^2 + y_4^2) \\
&= (x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4)^2 + (x_1 y_2 - x_2 y_1 + x_3 y_4 - x_4 y_3)^2 \\
&+ (x_1 y_3 - x_3 y_1 + x_2 y_4 - x_4 y_2)^2 + (x_1 y_4 - x_4 y_1 + x_2 y_3 - x_3 y_2)^2.
\end{aligned}$$

Proof. Direct verification. □

The identity shows that if each of two numbers $a, b \in \mathbf{N}_0$ is a sum of four squares then so is their product ab.

Arithmetic proof of the four squares theorem. By the identity it suffices to prove the theorem only for primes $n = p$. Since $2 = 1^2 + 1^2 + 0^2 + 0^2$, we may assume that $p > 2$. The argument of the above lemma (applied to numbers a^2 and $-a^2$) shows that there are $a, b \in \{0, 1, \ldots, (p-1)/2\}$ and $m \in \mathbf{N}$ such that

$$mp = a^2 + b^2 = a^2 + b^2 + 0^2 + 0^2 = x_1^2 + x_2^2 + x_3^2 + x_4^2.$$

Clearly, $1 \le m < p/4 + p/4 = p/2$. If $m = 1$, we are done. Let us assume that $1 < m < p/2$. We shall find a smaller nonzero multiple of p that is a sum of four squares. Repeating this reduction, in the end we express p as a sum of four squares.

There are unique four integers y_i such that

$$y_i \equiv x_i \pmod m \quad \text{and} \quad -\frac{m}{2} < y_i \leq \frac{m}{2}.$$

Since $\sum_1^4 y_i^2 \equiv \sum_1^4 x_i^2 \equiv 0$ modulo m,

$$y_1^2 + y_2^2 + y_3^2 + y_4^2 = nm$$

for some $n \in \mathbf{N}_0$. We have $0 \leq n \leq m$. We show that equalities here are impossible. If $n = 0$ then all y_i are zero and every x_i is divisible by m. But then $\sum_1^4 x_i^2$ is a multiple of m^2 and from $\sum_1^4 x_i^2 = mp$ we get that m divides p which is not possible. If $n = m$ then every y_i equals $m/2$ and hence $x_i^2 \equiv m^2/4$ modulo m^2 (squaring $x_i = m/2 + r_i m$ gives $x_i^2 = m^2/4 + r_i m^2 + r_i^2 m^2$). Again, $\sum_1^4 x_i^2$ is a multiple of m^2 which is not possible. Thus

$$0 < n < m.$$

Multiplying equations $mp = \sum_1^4 x_i^2$ and $nm = \sum_1^4 y_i^2$ we get

$$nm^2p = (x_1^2 + x_2^2 + x_3^2 + x_4^2)(y_1^2 + y_2^2 + y_3^2 + y_4^2) = z_1^2 + z_2^2 + z_3^2 + z_4^2$$

where z_i are expressed in terms of x_i and y_i in Euler's four squares identity. These expressions and congruences $y_i \equiv x_i$, $\sum_1^4 x_i^2 \equiv 0$ modulo m imply that every z_i is a multiple of m: $z_i = mu_i$, $u_i \in \mathbf{Z}$, $i = 1, \ldots, 4$. Dividing the displayed equality by m^2 we get

$$np = u_1^2 + u_2^2 + u_3^2 + u_4^2.$$

Since $1 \leq n < m$, the promise is fulfilled. $\square$

Let $r_4(n)$ denote the number of solutions $(x_1, x_2, x_3, x_4) \in \mathbf{Z}^4$ of the equation

$$n = x_1^2 + x_2^2 + x_3^2 + x_4^2.$$

(Solutions differing only by signs or by order of summands are still counted as different.) In 1829, Carl Jacobi (1804–1851) proved a miraculous identity: If $n \geq 1$ is an integer then

$$r_4(n) = 8 \sum_{d \mid n,\ 4 \nmid d} d.$$

For example, for $n = 12$ we get $r_4(12) = 8(1+2+3+6) = 96$ which agrees with the representations $12 = 2^2 + 2^2 + 2^2 + 0^2$ contributing $4 \cdot 2^3 = 32$ solutions and $12 = 3^2 + 1^2 + 1^2 + 1^2$ contributing $4 \cdot 2^4 = 64$ solutions. Euler's four squares identity is trivial to verify, once it is stated, but Jacobi's identity is difficult to prove. Indeed, its simple consequence is Lagrange's theorem saying that $r_4(n) > 0$ for every $n \in \mathbf{N}$: the sum in the identity always contains summand 1 and thus in fact always $r_4(n) \ge 8$.

3.3 The circle problem and the divisor problem

Similarly, let $r_2(n)$ denote the number of solutions $(x_1, x_2) \in \mathbf{Z}^2$ of

$$n = x_1^2 + x_2^2,$$

that is, $r_2(n)$ counts expressions of n as a sum of two squares. What can be said about the asymptotics of the summatory function

$$\sum_{n \le x} r_2(n)$$

when $x \to \infty$? A geometric insight of Carl Friedrich Gauss (1777-1855) provided an answer.

Theorem (Gauss, cca 1800). *For $x \to \infty$,*

$$\sum_{n \le x} r_2(n) = \pi x + O(x^{1/2}).$$

Proof. The summand $r_2(n)$ is the number of the lattice points lying on the origin-centered circle with radius $\sqrt{n}$ and therefore the sum equals the number of the lattice points in K where K is the disc with center in the origin and radius $\sqrt{x}$. Let C be the set of all unit squares $[a - \frac{1}{2}, a + \frac{1}{2}] \times [b - \frac{1}{2}, b + \frac{1}{2}]$ which have center $(a, b) \in \mathbf{Z}^2$ and intersect K. Let B be those of them which have centers in K and A be those which lie completely in K. Then

$$A \subset B \subset C,\ |B| = \sum_{n \le x} r_2(n), \quad \text{and} \quad |A| \le \mathrm{Vol}(K) = \pi x \le |C|.$$

From this, $|A| - |B| \le \mathrm{Vol}(K) - \sum_{n\le x} r_2(n) \le |C| - |B|$ and

$$\left|\sum_{n\le x} r_2(n) - \pi x\right| \le |C| - |A| = |C\backslash A|.$$

The set $C\backslash A$ consists of the squares which intersect both K and the complement of K. These squares all lie in the annulus

$$L = \{z \in \mathbf{R}^2 : \sqrt{x} - \sqrt{2} \le |z| \le \sqrt{x} + \sqrt{2}\}$$

because they intersect the boundary circle of K and unit square has diameter $\sqrt{2}$. Thus

$$|C\backslash A| \le \mathrm{Vol}(L) = \pi((\sqrt{x}+\sqrt{2})^2 - (\sqrt{x}-\sqrt{2})^2) = 4\sqrt{2}\pi\sqrt{x}$$

and the bound on the error follows. □

Dirichlet derived similar asymptotics for the sum

$$\sum_{n\le x} d(n)$$

where $d(n)$ is the number of divisors of n, which is the number of solutions $(a,b) \in \mathbf{N}^2$ of the equation $ab = n$.

Theorem (Dirichlet, 1849). *For $x \to \infty$,*

$$\sum_{n\le x} d(n) = x\log x + (2\gamma - 1)x + O(x^{1/2}).$$

Here $\gamma = 0.57722\ldots$ is the *Euler–Mascheroni constant* that is defined in the following lemma.

Lemma. *For $x \to \infty$,*

$$\sum_{n\le x} \frac{1}{n} = \log x + \gamma + O(x^{-1})$$

where $\gamma > 0$ is a constant.

Proof. From ($n \in \mathbf{N}$)

$$\int_n^{n+1} \frac{dt}{t} = \log(1+1/n) = \frac{1}{n} + z(n),$$

where $z(n) = -n^{-2}/2 + n^{-3}/3 - \cdots = O(n^{-2})$ and is negative for big n, we get ($N \in \mathbf{N}$)

$$\begin{aligned} \sum_{n=1}^{N} \frac{1}{n} &= \sum_{n=1}^{N} \left(\int_n^{n+1} \frac{dt}{t} - z(n) \right) = \int_1^{N+1} \frac{dt}{t} - \sum_{n=1}^{\infty} z(n) + \sum_{n>N} z(n) \\ &= \log N + \gamma + O(1/N), \end{aligned}$$

because $\log(N+1) = \log N + O(1/N)$, $\sum_{n=1}^{\infty} z(n)$ converges (to a sum γ), and $\sum_{n>N} z(n) = O(1/N)$ by a simple integral estimate. The asymptotics still holds after replacing $N \in \mathbf{N}$ by $x \in \mathbf{R}$, $x > 0$. □

It is not known whether γ is an irrational number or not.

Proof of Dirichlet's theorem. The summand $d(n)$ is the number of the lattice points on the branch of the hyperbola $\{(a,b) \in \mathbf{R}^2 : \ ab = n, a > 0\}$. The sum is therefore equal to the number of the lattice points in the plane region

$$A = \{(a,b) \in \mathbf{R}^2 : \ ab \le x, \ a, b > 0\}.$$

We write

$$A = B \cup C$$

where

$$B = \{(a,b) \in A : \ a \le \sqrt{x}\} \ \text{ and } \ C = \{(a,b) \in A : \ b \le \sqrt{x}\}.$$

The intersection $B \cap C$ is the plane square $K = (0, \sqrt{x}\,]^2$. Regions B and C are reflections of one another along the line $a = b$ and therefore contain the same numbers of lattice points. Thus

$$\sum_{n \le x} d(n) = 2|\mathbf{Z}^2 \cap B| - |\mathbf{Z}^2 \cap K|$$

which equals

$$2 \sum_{n \le \sqrt{x}} \left\lfloor \frac{x}{n} \right\rfloor - \lfloor \sqrt{x} \rfloor^2.$$

Omitting floors in the sum causes total error $O(x^{1/2})$ because each of the $\lfloor\sqrt{x}\rfloor$ summands contributes error at most 1. Using the lemma we have

$$\begin{aligned} 2\sum_{n\le\sqrt{x}}\left\lfloor\frac{x}{n}\right\rfloor &= 2x\sum_{n\le\sqrt{x}}\frac{1}{n}+O(x^{1/2}) \\ &= 2x(\log\sqrt{x}+\gamma+O(x^{-1/2}))+O(x^{1/2}) \\ &= x\log x+2\gamma x+O(x^{1/2}). \end{aligned}$$

Subtracting $\lfloor\sqrt{x}\rfloor^2 = x - 2\sqrt{x}\{\sqrt{x}\}+\{\sqrt{x}\}^2 = x+O(x^{1/2})$ we obtain the asymptotics. $\square$

The *circle problem*, resp. the *divisor problem*, asks to find the infimum of all exponents $\alpha > 0$ such that $O(x^\alpha)$ is a valid estimate of the error in the asymptotics of $\sum_{n\le x} r_2(n)$, resp. of $\sum_{n\le x} d(n)$. The previous two theorems show that both infima are $\le \frac{1}{2}$. Godfrey Hardy (1877-1947) proved that they are $\ge \frac{1}{4}$. Martin Huxley established in 2003 the current record in the upper bound by proving that both infima are $\le \frac{131}{416}$ $(= 0.3149038\ldots)$. It is conjectured that they are equal to $\frac{1}{4}$.

3.4 Remarks

The exposition follows mostly Hlawka, Schoißengeier and Taschner [10]. Huxley's record exponent is obtained in [11].

Chapter 4

Prime numbers

We shall discuss properties of the multiplicative "atoms" in the world of integers, the prime numbers. We give four proofs of the infinititude of their number and prove the classical result of Chebyshev bounding the number of primes up to x from below and from above by constant multiples of the function $x/\log x$. Then we derive the asymptotic relations for $\sum_{p\le x} \log p/p$, $\sum_{p\le x} 1/p$ and $\prod_{p\le x}(1-1/p)$ due to Mertens. We show that natural numbers n have in average $\log\log n$ prime factors, no matter if we count them with or without multiplicity, and that in fact almost all numbers n have $\log\log n$ prime factors (the Hardy–Ramanujan theorem).

4.1 Four proofs of the infinititude of primes

Primes

$$2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, \ldots$$

are those natural numbers n bigger than 1 which have only the trivial divisors 1 and n. We shall denote them by the letters p and q. The value of the *prime counting function* $\pi(x)$ is for $x \in \mathbf{R}$ defined as the number of primes not exceeding x:

$$\pi(x) = \#\{p : \ p \le x\}.$$

For example, $\pi(-3) = \pi(1.9) = 0$ and $\pi(18) = 7$. The importance of primes stems from the fact that every $n \in \mathbf{N}$ has a unique prime factorization

$$n = p_1^{a_1} p_2^{a_2} \ldots p_r^{a_r}$$

where $p_1 < p_2 < \ldots < p_r$ are distinct primes and $a_i \in \mathbf{N}$ — this is so called *Fundamental Theorem of Arithmetic* (FTA). Already Euclid could prove that there are infinitely many primes.

Theorem (Euclid, cca 500 B.C.). *There exist infinitely many prime numbers.*

We shall present four proofs.

1. Euclid's proof. First note that every $n \in \mathbf{N}$ bigger than 1 is divisible by a prime (take the minimal divisor of n bigger than 1). If there were only finitely many primes $p_1, p_2, \ldots, p_r$, consider a prime divisor p of the number

$$p_1 p_2 \ldots p_r + 1.$$

We must have $p = p_i$ for some i. Thus p_i divides also 1, which is a contradiction. $\square$

2. Goldbach's proof. Consider the recurrent sequence $(G_n)_{n\geq 0}$ given by $G_0 = 2$ and

$$G_n = G_0 G_1 \ldots G_{n-1} + 1 \quad \text{for} \quad n \in \mathbf{N}.$$

Hence $G_0 = 2, G_1 = 3, G_2 = 7, G_3 = 43$ and so on. Clearly, if $m < n$ then G_m, G_n are coprime (because G_m divides $G_n - 1$). Selecting an arbitrary prime divisor of each G_n, we get infinitely many primes that must be all distinct. $\square$

This proof is usually cast with a slightly more complicated recurrent sequence $(F_n)_{n\geq 0}$,

$$F_0 = 3 \quad \text{and} \quad F_n = F_0 F_1 \ldots F_{n-1} + 2 \quad \text{for} \quad n \in \mathbf{N}.$$

Now

$$F_0 = 3,\ F_1 = 5,\ F_2 = 17,\ F_3 = 257,\ F_4 = 65537,\ \ldots\ .$$

Again, the numbers F_n are pairwise coprime and produce infinitely many primes. The advantage of F_n over G_n is that relation $F_{n+1} = (F_n - 2)F_n + 2 = (F_n - 1)^2 + 1$ and induction give

$$F_n = 2^{2^n} + 1.$$

So F_n, unlike G_n, can be defined by an explicit formula.

Numbers F_n are called *Fermat numbers* and in the case that F_n is a prime number it is called *Fermat prime*. One of the most interesting properties of Fermat primes is the following result.

Theorem (Gauss, 1797; Wantzel, 1832). *Regular plane k-gon can be constructed by ruler and compass if and only if $k = 2^r p_1 p_2 \ldots p_s$ where $r \in \mathbf{N}_0$ and p_i are distinct Fermat primes.*

This beautiful result is often attributed completely to Gauss but although Gauss stated it, he proved only the implication $\Leftarrow$. The implication $\Rightarrow$ was proved by Pierre Wantzel (1814–1848). [1] The only Fermat primes known to date are the five listed above, F_n with $0 \le n \le 4$. It is not known if any F_n for $n > 4$ is prime.

3. Euler's proof, 1st variant. It is based on the *Euler identity*

$$\prod_p \frac{1}{1 - 1/p^s} = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

that holds for every real $s > 1$. We will not give a rigorous proof which is not difficult. Note that on the left the p-th factor is the sum of the geometric series $1 + 1/p^s + 1/p^{2s} + \cdots$. Multiplying the factors (and disregarding convergence matters) we get, due to the FTA, the series on the right.

Suppose that there are only finitely many primes $p_1, \ldots, p_r$ and see what happens with the identity if $s \to 1^+$. The left side has the finite limit

$$\frac{1}{(1 - p_1^{-1})(1 - p_2^{-1}) \ldots (1 - p_r^{-1})}.$$

However, the right side goes to $+\infty$ because its partial sums more and more closely approximate partial sums of the divergent harmonic series $\sum 1/n$. We have a contradiction and there must be infinitely many primes. $\square$

2nd variant of Euler's proof. For two infinite series $A = \sum_{n \ge 0} a_n$ and $B = \sum_{n \ge 0} b_n$ of nonnegative real numbers we define the product series $C = AB$ as

$$C = \sum_{n \ge 0} c_n = a_0 b_0 + (a_0 b_1 + a_1 b_0) + (a_0 b_2 + a_1 b_1 + a_2 b_0) + \cdots,$$

[1] Such pattern of (mis)attribution of results to people is an example of so called Matthew effect (Matthew 25:29).

that is, $c_n = \sum_{k=0}^{n} a_k b_{n-k}$. Clearly, if A and B converge then so does C because

$$\sum_{k=0}^{n} c_k = \sum_{k=0}^{n}\sum_{i=0}^{k} a_i b_{k-i} \le \sum_{k=0}^{n} a_k \sum_{l=0}^{n} b_l \le \sum_{k=0}^{\infty} a_k \sum_{l=0}^{\infty} b_l.$$

For a finite product of more than two series we proceed similarly (or by induction, if you wish).

Now suppose that there are only finitely many primes, $p_1, p_2, \ldots, p_k$, and for $i = 1, 2, \ldots, k$ and $s > 0$ consider the geometric series

$$S_i(s) = \sum_{k=0}^{\infty} \frac{1}{p_i^{sk}}.$$

It converges for every $s > 0$ and, by the just proved lemma, so does the product series

$$S_1(s)S_2(s)\ldots S_k(s) = \prod_{i=1}^{k}\sum_{l=0}^{\infty} \frac{1}{p_i^{sl}} = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

where the last equality follows from the definition of product series and from the FTA. However, this is a contradiction because the last series is divergent for $0 < s \le 1$. □

4. Erdős' proof. Every $n \in \mathbf{N}$ can be written as $n = r^2 s$ where $r, s \in \mathbf{N}$ and s is squarefree (the representation is unique but we will not need this). If $n \le N$, $N \in \mathbf{N}$, then r attains at most $\sqrt{N}$ values (because $r^2 \le n \le N$) and s attains at most $2^{\pi(N)}$ values (the possible values of s correspond to the subsets of the $\pi(N)$-element set of the primes not exceeding N because $s = p_1 p_2 \ldots p_r$ for distinct primes $p_i \le N$). Necessarily $\sqrt{N} \cdot 2^{\pi(N)} \ge N$ because else there would not be enough representations $r^2 s$ for all $n \le N$. Thus

$$\pi(N) \ge \tfrac{1}{2}\log_2 N$$

and $\pi(x) \to \infty$ as $x \to \infty$. □

4.2 Theorems of Chebyshev and Mertens

In the last proof, invented by Paul (Pál) Erdős (1913–1996), a lower bound on the prime counting function $\pi(x)$ was derived. It is a rather weak one, though, as the next estimate shows. It is due to Pafnuty Lvovich Chebyshev (1821–1894) [2] but the proof presented here, based on the properties of the middle binomial coefficient $\binom{2n}{n}$, is due to Erdős.

Theorem (Chebyshev, 1850). *For every $x \geq 2$,*

$$\frac{c_1 x}{\log x} < \pi(x) < \frac{c_2 x}{\log x}$$

where $0 < c_1 < c_2$ are constants.

Proof (Erdős, 1936). Let $n \in \mathbf{N}$. We have the estimate

$$\frac{4^n}{2n+1} \leq \binom{2n}{n} \leq 4^n$$

because $4^n = (1+1)^{2n} = \sum_{k=0}^{2n} \binom{2n}{k}$ and $\binom{2n}{n}$ is the largest of the $2n+1$ binomial coefficients in the sum.

Chebyshev's bounds are obtained by combining this with another estimate of the middle binomial coefficient:

$$\prod_{n<p\leq 2n} p \leq \binom{2n}{n} \leq (2n)^{\pi(2n)}.$$

The lower bound in this estimate follows from the fact that in

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2}$$

the numerator contributes all primes from the interval $(n, 2n]$ and they cannot be canceled because the primes in the denominator are all $\leq n$. To prove the upper bound, we estimate p^a, the highest power of a prime p dividing $\binom{2n}{n}$. We claim that

$$a = \sum_{i=1}^{\infty} \lfloor 2n/p^i \rfloor - 2\lfloor n/p^i \rfloor = \sum_{i,\ p^i \leq 2n} \lfloor 2n/p^i \rfloor - 2\lfloor n/p^i \rfloor.$$

This is a corollary of the formula $b = \sum_{i \geq 1} \lfloor m/p^i \rfloor$ for the highest exponent $b \in \mathbf{N}$ such that p^b divides $m! = 1 \cdot 2 \cdot \ldots \cdot m$—the i-th summand counts

[2] Pafnutij Lvovič Čebyšev, Пафнутий Львович Чебышев.

multiples of p^i among the numbers $1, 2, \ldots, m$. But $0 \le \lfloor 2\alpha \rfloor - 2\lfloor \alpha \rfloor \le 1$ for every $\alpha \in \mathbf{R}$. Hence

$$a \le \sum_{i,\ p^i \le 2n} 1 \quad \text{and therefore} \quad p^a \le 2n.$$

Surprisingly, the prime powers in the factorization of $\binom{2n}{n}$ are only as big as if we factorized $2n$. Every prime in the factorization is $\le 2n$. Thus we have $\le \pi(2n)$ prime powers of size $\le 2n$ and their product is $\le (2n)^{\pi(2n)}$.

Combining the two estimates of $\binom{2n}{n}$ we get

$$(2n)^{\pi(2n)} \ge \frac{4^n}{2n+1}.$$

Taking logarithms, we obtain

$$\pi(2n) \ge \frac{2n \cdot \log 2}{\log(2n)} - \frac{\log(2n+1)}{\log(2n)} > \frac{2n \cdot \log 2}{\log(2n)} - 2$$

and (for $n \ge 2$)

$$\pi(2n-1) = \pi(2n) > \frac{2n \cdot \log 2}{\log(2n)} - 2 \ge \frac{(2n-1) \cdot \log 2}{\log(2n-1)} - 2,$$

which is the lower bound of Chebyshev.

Combining the two estimates of $\binom{2n}{n}$ in the other way and taking logarithms we get

$$\sum_{n<p\le 2n} \log p \le n \log 4.$$

To get an upper bound for the sum over all $p \le x$, $x \ge 2$, we take the largest $m \in \mathbf{N}$ such that $2^m \le x$. Then $2^{m+1} > x$ and

$$\begin{aligned} \sum_{p\le x} \log p &\le \sum_{k=0}^{m} \left(\sum_{2^k<p\le 2^{k+1}} \log p \right) \le (2^0 + 2^1 + \cdots + 2^m) \log 4 < 2^{m+1} \log 4 \\ &\le (2 \log 4)x. \end{aligned}$$

From this we have

$$\begin{aligned} (2\log 4)x &\ge \sum_{p\le x} \log p \ge \sum_{\sqrt{x}<p\le x} \log p \ge (\pi(x) - \pi(\sqrt{x})) \log \sqrt{x} \\ &\ge (\pi(x) - \sqrt{x}) \log \sqrt{x} \end{aligned}$$

and therefore

$$\pi(x) \le \frac{(4\log 4)x}{\log x} + \sqrt{x},$$

which is the upper bound of Chebyshev. $\square$

The precise asymptotics of $\pi(x)$, the *Prime Number Theorem*, was proved independently by Jacques Hadamard (1865–1963) and Charles de La Valée Poussin (1866–1962).

Theorem (Hadamard, 1896; de La Valée Poussin, 1896). *For $x \to \infty$,*

$$\pi(x) = (1 + o(1))\frac{x}{\log x}.$$

We will not give the proof here.

The *von Mangoldt function* $\Lambda : \mathbf{N} \to \mathbf{R}$, named after Hans von Mangoldt (1854-1925), is defined by

$$\Lambda(n) = \begin{cases} \log p & \text{if } n = p^r \\ 0 & \text{else.} \end{cases}$$

Lemma. *For $x \to \infty$,*

$$\sum_{n \le x} \Lambda(n)\lfloor x/n \rfloor = \log(\lfloor x \rfloor!) = x\log x + O(x).$$

Proof. We write the summand $\lfloor x/n \rfloor$ as the number of multiples of n not exceeding x and then change the order of summation:

$$\sum_{n \le x} \Lambda(n) \sum_{m \le x, n|m} 1 = \sum_{m \le x} \sum_{n|m} \Lambda(n) = \sum_{m \le x} \log m = \log(\lfloor x \rfloor!),$$

because for $m = p_1^{a_1} \dots p_r^{a_r}$ the last inner sum is

$$a_1 \log p_1 + a_2 \log p_2 + \dots + a_r \log p_r = \log m.$$

We proved the first equality. The asymptotics of $\log(\lfloor x \rfloor!)$ follows from a simple integral estimate similar to the lemma on harmonic numbers. $\square$

Useful summation technique described in the following lemma is due to Niels Abel (1802–1829).

Lemma (Abel's summation). *If $(a_n)_{n\geq 1}$ is a sequence of real numbers with summatory function $A(x)=\sum_{n\leq x} a_n$ and $f(x)$ is a real function which has first derivative on the interval $(1-\varepsilon,\infty)$, then for every $x\geq 1$ we have the identity*

$$\sum_{n\leq x} a_n f(n) = A(x)f(x) - \int_1^x A(t)f'(t)\,dt.$$

If $r\geq 1$ and $f(x)$ has first derivative on the interval $(r-\varepsilon,\infty)$, then for every $x\geq r$ we have the identity

$$\sum_{r<n\leq x} a_n f(n) = A(x)f(x) - A(r)f(r) - \int_r^x A(t)f'(t)\,dt.$$

Proof. Let $N=\lfloor x\rfloor$. Note that $A(0)=0$ and $A(x)=A(N)$. Using the relation $a_n = A(n)-A(n-1)$, we transform the sum:

$$\begin{aligned}
\sum_{n\leq x} a_n f(n) = \sum_{n=1}^{N} a_n f(n) &= \sum_{n=1}^{N} (A(n)-A(n-1))f(n)\\
&= \sum_{n=1}^{N-1} A(n)(f(n)-f(n+1)) + A(N)f(N)\\
&= A(N)f(N) - \sum_{n=1}^{N-1} A(n)\int_n^{n+1} f'(t)\,dt\\
&= A(N)f(N) - \sum_{n=1}^{N-1} \int_n^{n+1} A(t)f'(t)\,dt\\
&= A(N)f(N) - \int_1^{N} A(t)f'(t)\,dt.
\end{aligned}$$

Since

$$\begin{aligned}
A(N)f(N) &= A(x)f(x) - A(N)(f(x)-f(N)) \text{ and}\\
\int_1^N A(t)f'(t)\,dt &= \int_1^x A(t)f'(t)\,dt - \int_N^x A(t)f'(t)\,dt\\
&= \int_1^x A(t)f'(t)\,dt - A(N)(f(x)-f(N)),
\end{aligned}$$

the first identity follows. To obtain the second identity, extend $f(x)$ arbitrarily so that it is defined and has first derivative on $(1-\varepsilon, \infty)$ and subtract the expression given by the first identity for $\sum_{n\le r} a_n f(n)$ from that for $\sum_{n\le x} a_n f(n)$. $\square$

It is difficult to find the asymptotics of the sum $\pi(x) = \sum_{p\le x} 1$. But if the summand goes reasonably to zero, precise asymptotics can be obtained by simple means. Classical result of this type is due to Franz Mertens (1840–1927).

Theorem (Mertens, 1874). *As $x \to \infty$, we have the following three asymptotics.*

$$\begin{aligned}
1.\quad \sum_{p\le x} \frac{\log p}{p} &= \log x + O(1),\\
2.\quad \sum_{p\le x} \frac{1}{p} &= \log\log x + c + O(1/\log x),\\
3.\quad \prod_{p\le x}\left(1-\frac{1}{p}\right) &= \frac{d}{\log x}(1+O(1/\log x))
\end{aligned}$$

where c, d are constants (of course, $d > 0$).

Proof. 1. By the lemma on $\Lambda(n)$ and the definition of $\Lambda(n)$,

$$\begin{aligned}
x\log x + O(x) &= \sum_{n\le x} \Lambda(n)\lfloor x/n\rfloor\\
&= x\sum_{p\le x}(\log p)/p + O(x) + \sum_{p^r\le x, r\ge 2}(\log p)\lfloor x/p^r\rfloor
\end{aligned}$$

(omitting floors in the first sum over primes not exceeding x creates an error of size at most $(\log x)O(x/\log x) = O(x)$ because $|\Lambda(n)| = \log p \le \log x$ and by Chebyshev's theorem the sum has $O(x/\log x)$ summands). The second sum is at most

$$x\sum_{n,r\ge 2}\frac{\log n}{n^r} \le x(1+1/2+1/4+1/8+\cdots)\sum_{n=2}^{\infty}\frac{\log n}{n^2} = O(x).$$

Dividing by x we obtain the first formula of Mertens.

2. To prove the second formula we set $f(x) = 1/\log x$ and

$$a_n = \begin{cases} \dfrac{\log p}{p} & \text{if } n = p \\ \\ 0 & \text{else.} \end{cases}$$

Note that the summatory function $A(x)$ is the sum in the first Mertens' formula. We have

$$\sum_{p \le x} \frac{1}{p} = \sum_{p \le x} \frac{\log p}{p} \cdot \frac{1}{\log p} = \sum_{1.5 < n \le x} a_n f(n).$$

Using Abel's summation and the asymptotics in 1 (we denote the error term in it as $z(x)$), we get

$$\begin{aligned} \sum_{p \le x} \frac{1}{p} &= A(x)f(x) - A(1.5)f(1.5) - \int_{1.5}^{x} A(t)f'(t)\,dt \\ &= 1 + O(1/\log x) + \int_{1.5}^{x} \frac{\log t + z(t)}{t \log^2 t}\,dt \\ &= 1 + O(1/\log x) + \int_{1.5}^{x} \frac{dt}{t \log t} + \int_{1.5}^{\infty} \frac{z(t)\,dt}{t \log^2 t} - \int_{x}^{\infty} \frac{z(t)\,dt}{t \log^2 t}. \end{aligned}$$

Because $(\log \log t)' = 1/(t \log t)$, $(-1/\log t)' = 1/(t \log^2 t)$ and $z(t) = O(1)$, the first integral evaluates as $\log \log x - \log \log 1.5$, the second integral converges to a constant c, and the third integral is $O(1/\log x)$. Thus

$$\sum_{p \le x} \frac{1}{p} = \log \log x + 1 - \log \log 1.5 + c + O(1/\log x)$$

and the second formula of Mertens follows.

3. Taking logarithm of the product and using expansion $\log(1-x) = -\sum_{n \ge 1} x^n/n$ we get

$$\begin{aligned} \sum_{p \le x} \log(1 - 1/p) &= -\sum_{p \le x} \sum_{k \ge 1} 1/(kp^k) = -\sum_{p \le x} \frac{1}{p} - \sum_{p \le x, k > 1} \frac{1}{kp^k} \\ &= -\sum_{p \le x} \frac{1}{p} - \sum_{k=2}^{\infty} \sum_{p} \frac{1}{kp^k} + \sum_{k=2}^{\infty} \sum_{p > x} \frac{1}{kp^k}. \end{aligned}$$

The first term is $-\log\log x + c_1 + O(1/\log x)$ by part 2. The second sum converges to $c_2 > 0$ (in fact, $c_2 \le \sum n^{-2} = \pi^2/6$). The third sum is at most

$$\sum_{k=2}^{\infty}\sum_{n>x}\frac{1}{n^k} \le \sum_{k=2}^{\infty}\int_{x-1}^{\infty}\frac{dt}{t^k} = \sum_{k=2}^{\infty}\frac{1}{(k-1)(x-1)^{k-1}} = O(1/x),$$

which is absorbed in $O(1/\log x)$. So

$$\sum_{p\le x}\log(1-1/p) = -\log\log x + c_1 - c_2 + O(1/\log x)$$

and

$$\prod_{p\le x}\left(1-\frac{1}{p}\right) = \frac{\mathrm{e}^{c_1-c_2}}{\log x}\cdot \mathrm{e}^{O(1/\log x)} = \frac{\mathrm{e}^{c_1-c_2}}{\log x}(1+O(1/\log x)),$$

which is the third formula of Mertens. □

In fact, the constant in the third formula is $d = \mathrm{e}^{\gamma}$ where γ is the Euler–Mascheroni constant.

4.3 Estimates of the functions $\omega(n)$ and $\Omega(n)$

The functions ω and Ω measure the complexity of the prime factorization $n = p_1^{a_1}p_2^{a_2}\dots p_r^{a_r}$ of $n \in \mathbf{N}$ and are defined by

$$\omega(n) = r \quad\text{and}\quad \Omega(n) = a_1 + a_2 + \cdots + a_r.$$

So $\omega(n)$ is the number of prime factors of n and $\Omega(n)$ is their number including the multiplicities. For example, $\omega(12) = 2$ and $\Omega(12) = 3$.

Lemma. *For $x \to \infty$,*

$$\begin{aligned}\sum_{n\le x}\omega(n) &= x\log\log x + c_1x + O(x/\log x) \text{ and}\\ \sum_{n\le x}\Omega(n) &= x\log\log x + c_2x + O(x/\log x)\end{aligned}$$

where c_1, c_2 are constants.

Proof. We write $\omega(n) = \sum_{p|n} 1$ and change order of summation. The first sum then equals

$$\begin{aligned}\sum_{n\le x}\omega(n) = \sum_{p\le x}\sum_{n\le x,\ p|n} 1 &= \sum_{p\le x}\lfloor x/p\rfloor = x\sum_{p\le x} 1/p + O(x/\log x)\\ &= x\log\log x + c_1 x + O(x/\log x).\end{aligned}$$

(Omitting floors we make an error E, where $|E| \le \pi(x) = O(x/\log x)$ by the upper bound of Chebyshev. Then we use the second formula of Mertens.)

In the second sum we write similarly $\Omega(n) = \sum_{p^m|n} 1$ and change order of summation:

$$\sum_{n\le x}\Omega(n) = \sum_{p^m\le x}\lfloor x/p^m\rfloor.$$

We need to estimate the surplus sum

$$A(x) = \sum_{n\le x}(\Omega(n) - \omega(n)) = \sum_p\sum_{m\ge 2}\lfloor x/p^m\rfloor.$$

On the one hand

$$A(x) \le x\sum_p\sum_{m\ge 2} p^{-m} = x\sum_p \frac{1}{p(p-1)}.$$

On the other hand

$$A(x) \ge \sum_{p\le\sqrt{x}}\ \sum_{2\le m\le \log x/\log p}(xp^{-m} - 1).$$

The inner sum equals

$$x\sum_{m\ge 2} p^{-m} - x\sum_{m>\log x/\log p} p^{-m} + O(\log x) = \frac{x}{p(p-1)} - c + O(\log x)$$

where $0 < c \le 2$. Thus

$$\begin{aligned}A(x) &\ge x\sum_p\frac{1}{p(p-1)} - x\sum_{p>\sqrt{x}}\frac{1}{p(p-1)} + O(\sqrt{x}\log x)\\ &= x\sum_p\frac{1}{p(p-1)} + O(\sqrt{x}\log x)\end{aligned}$$

because

$$x \sum_{p>\sqrt{x}} \frac{1}{p(p-1)} \le 2x \sum_{n>\sqrt{x}} \frac{1}{n^2} \le 2x \int_{\sqrt{x}-1}^{\infty} t^{-2}\, dt = O(\sqrt{x}).$$

So $A(x) = cx + O(\sqrt{x}\log x)$, where $c = \sum_p \frac{1}{p(p-1)}$, and the second asymptotics follows. $\square$

In average the numbers $n \le x$ have $\frac{1}{x}\sum_{n\le x}\omega(n) \sim \log\log x$ prime divisors and the same holds for the average value of $\Omega(n)$. Can it be inferred that for almost all numbers $n \le x$ we have $\omega(n), \Omega(n) \approx \log\log x$? Yes. This was proved by Godfrey Harold Hardy (1877–1947) and Srinivasa Ramanujan (1887–1920). The proof that we present here is due to Paul Turán (1910–1976).

Theorem (Hardy and Ramanujan, 1917). *For every $\varepsilon > 0$ there is an $x_0 = x_0(\varepsilon)$ such that for every $x > x_0$,*

$$\#\{n \le x :\ |\omega(n) - \log\log x| < \varepsilon \log\log x\} > (1-\varepsilon)x.$$

This remains true when $\omega(n)$ is replaced by $\Omega(n)$.

Before starting with the proof we justify the last claim. Suppose that the theorem holds with $\omega(n)$ and fix $\varepsilon > 0$. Clearly $\omega(n) \le \Omega(n)$ for every n. By the proof of the previous lemma, $\sum_{n\le x}(\Omega(n)-\omega(n)) = cx+O(\sqrt{x}\log x)$ with a constant $c > 0$. Thus for any fixed $k \in \mathbf{N}$ and sufficiently large x, the inequality $\Omega(n) - \omega(n) > 2k$ may hold only for at most cx/k numbers $n \le x$. Hence $\omega(n) \le \Omega(n) \le \omega(n) + 2k$ holds for at least $(1 - c/k)x$ numbers $n \le x$, if $x > x_1(k)$. We fix $k \in \mathbf{N}$ so that $c/k < \varepsilon$. Then for every big x satisfying $x > x_0(\varepsilon)$, $x > x_1(k)$ and $2k < \varepsilon \log\log x$ we have $|\Omega(n) - \log\log x| \le |\Omega(n) - \omega(n)| + |\omega(n) - \log\log x| < 2\varepsilon \log\log x$ for at least $(1-2\varepsilon)x$ numbers $n \le x$. Similarly it is easy to see that in the theorem we may write $\log\log n$ instead of $\log\log x$.

Proof (Turán, 1937). First, we derive an asymptotics for $\sum_{n\le x}\omega(n)^2$ as $x \to \infty$. We represent the summand as $\omega(n)^2 = (\sum_{p|n} 1)(\sum_{q|n} 1) = \sum_{p|n,q|n} 1$ (because $1^2 = 1$). Interchanging summation order and putting apart the pairs p, q with $p = q$ we get

$$\sum_{n\le x}\omega(n)^2 \ = \ \sum_{p\ne q}\ \sum_{p|n,q|n,n\le x} 1 + \sum_{p=q}\ \sum_{p|n,q|n,n\le x} 1$$

$$= \sum_{pq\le x,\ p\ne q} \lfloor x/pq\rfloor + \sum_{p=q\le x} \lfloor x/p\rfloor.$$

The first sum has at most $2x$ summands and omitting floors in it creates an error $O(x)$. Omitting the condition $p \ne q$ increases it by at most $\sum_{n\ge 1} x/n^2 = O(x)$. By the previous lemma and its proof, the second sum equals $\sum_{n\le x}\omega(n) = O(x\log\log x)$. Thus

$$\sum_{n\le x}\omega(n)^2 = x\sum_{pq\le x} 1/pq + O(x\log\log x).$$

It is easy to see that

$$\Bigg(\sum_{p\le\sqrt{x}} 1/p\Bigg)^2 \le \sum_{pq\le x} 1/pq \le \Bigg(\sum_{p\le x} 1/p\Bigg)^2.$$

The second formula of Mertens shows that both the first and the third sum are equal to $\log\log x + O(1)$. Thus

$$x\sum_{pq\le x} 1/pq = x(\log\log x + O(1))^2 = x(\log\log x)^2 + O(x\log\log x)$$

and we conclude that

$$\sum_{n\le x}\omega(n)^2 = x(\log\log x)^2 + O(x\log\log x).$$

As a corollary, using again the previous lemma, we obtain

$$\begin{aligned}\sum_{n\le x}(\omega(n)-\log\log x)^2 &= \sum_{n\le x}\omega(n)^2 - 2(\log\log x)\sum_{n\le x}\omega(n) + \sum_{n\le x}(\log\log x)^2\\ &= x(\log\log x)^2(1-2+1) + O(x\log\log x)\\ &= O(x\log\log x).\end{aligned}$$

Now we prove the theorem and in fact in a stronger form. Let $a(x)$ be any real function satisfying $a(x)\to\infty$ for $x\to\infty$. We show that for any $\varepsilon > 0$ there is an $x_0 = x_0(\varepsilon)$ such that if $x > x_0$,

$$\#\{n\le x:\ |\omega(n)-\log\log x| < a(x)(\log\log x)^{1/2}\} > (1-\varepsilon)x.$$

Suppose that this claim does not hold. Thus there is an $\varepsilon > 0$ and an infinite and to infinity going sequence $0 < x_1 < x_2 < \ldots$ such that for every i,

$$\#\{n \le x_i : \ |\omega(n) - \log\log x_i| \ge a(x_i)(\log\log x_i)^{1/2}\} \ge \varepsilon x_i.$$

But this implies

$$\sum_{n \le x_i} (\omega(n) - \log\log x_i)^2 \ge \varepsilon x_i \cdot a(x_i)^2 \log\log x_i,$$

which contradicts the asymptotics $\sum_{n\le x}(\omega(n) - \log\log x)^2 = O(x \log\log x)$ because $a(x_i)^2 \to \infty$ as $i \to \infty$. □

Dirichlet's theorem in chapter 3 implies that the average number of divisors $d(n)$ of the numbers $n \le x$ is $\approx \log x$. Can it be inferred that for almost all numbers $n \le x$ one has $d(n) \approx \log x$? (As in the previous theorem, it is irrelevant if we write here $\log x$ or $\log n$ because $\log x + \log \varepsilon < \log n \le \log x$ for $\varepsilon x < n \le x$.) No! Surprisingly, the correct typical value of $d(n)$ is considerably smaller:

$$d(n) \approx (\log n)^{\log 2} = (\log n)^{0.69314\ldots}.$$

This follows from the Hardy–Ramanujan theorem.

Corollary. *For every $\varepsilon > 0$ there is an $x_0 = x_0(\varepsilon)$ such that if $x > x_0$,*

$$\#\{n \le x : \ (\log n)^{\log 2 - \varepsilon} < d(n) < (\log n)^{\log 2 + \varepsilon}\} > (1 - \varepsilon)x.$$

Proof. Let $n \in \mathbf{N}$ and $n = p_1^{a_1} p_2^{a_2} \ldots p_r^{a_r}$. Then

$$d(n) = (a_1 + 1)(a_2 + 1) \ldots (a_r + 1).$$

Since $2^1 \le a + 1 \le 2^a$ for every $a \in \mathbf{N}$, we have

$$2^{\omega(n)} \le d(n) \le 2^{\Omega(n)}.$$

Fix $\varepsilon > 0$. By the theorem, for big x for more than $(1 - \varepsilon/2)x$ values of $n \le x$ the value $\omega(n)$ lies within $\varepsilon \log\log n$ to $\log\log n$, and the same holds for $\Omega(n)$. Hence for more than $(1 - \varepsilon)x$ values of $n \le x$ both $\omega(n)$ and

$\Omega(n)$ are, simultaneously, within $\varepsilon \log\log n$ to $\log\log n$. Using the estimate of $d(n)$ in terms of $\omega(n)$ and $\Omega(n)$ we see that for these n

$$(\log n)^{(1-\varepsilon)\log 2} = 2^{(1-\varepsilon)\log\log n} < d(n) < 2^{(1+\varepsilon)\log\log n} = (\log n)^{(1+\varepsilon)\log 2}.$$

$\square$

The second application of Hardy–Ramanujan's theorem — so called Erdős's multiplication table problem — concerns asymptotics of the counting function

$$A(x) = \#\{ab:\ a, b \in \mathbf{N}, a, b \le x\}.$$

We make, for $x \in \mathbf{N}$, an $x \times x$ multiplication table with rows and columns indexed with $1, 2, \ldots, x$. The number of cells is x^2 but what is the total number $A(x)$ of *distinct* products in the table? There are repetitions, for example 12 appears six times because $12 = 1 \cdot 12 = 12 \cdot 1 = 2 \cdot 6 = 6 \cdot 2 = 3 \cdot 4 = 4 \cdot 3$ but on the other hand every prime number appears only twice. Is the multiplicity of numbers in the table in average bounded, which means that $A(x) > cx^2$ for all $x \ge 1$ and some constant $c > 0$, or is it unbounded, which means that $A(x) = o(x^2)$?

Corollary. *For $x \to \infty$, $A(x) = o(x^2)$. Thus the average multiplicity of products in the multiplication table goes to infinity.*

Proof. We make use of the fact that $\Omega(n)$ is additive: $\Omega(mn) = \Omega(m) + \Omega(n)$. If a, b are two typical numbers $n \le x$ then, by the theorem, $\Omega(a), \Omega(b) \approx \log\log x$ and therefore

$$\Omega(ab) = \Omega(a) + \Omega(b) \approx 2\log\log x.$$

On the other hand, if c is a typical number $n \le x^2$, then

$$\Omega(c) \approx \log\log x^2 = \log\log x + \log 2,$$

which is asymptotically smaller than $2\log\log x$. It follows that the product ab of two typical numbers $n \le x$ is not a typical number $n \le x^2$ and hence we have only negligibly many products, $o(x^2)$.

More rigorous argument (demanded in the lecture). Fix ε with $0 < \varepsilon < 1/3$. Let $U(x)$ be the set of all $n \le x$ for which $|\Omega(n) - \log\log x| < \varepsilon \log\log x$, and $V(x)$ be the complement of $U(x)$ (in $\{1, 2, \ldots, x\}$). For $x > x_0$ we have $|U(x)| > (1-\varepsilon)x$ and $|V(x)| \le \varepsilon x$. Consider the pairs

(a, b), $a, b \le x$, and the products ab. The number of pairs with a or b in $V(x)$ is

$$2|U(x)| \cdot |V(x)| + |V(x)|^2 \le (2\varepsilon + \varepsilon^2)x^2$$

and this is an upper bound on the number of distinct products ab with a or b in $V(x)$. If both factors a, b are in $U(x)$ then

$$\Omega(ab) = \Omega(a) + \Omega(b) = 2 \log\log x + E(a, b)$$

where $|E(a, b)| \le 2\varepsilon \log\log x$. Hence, for big enough x, then ab must be in $V(x^2)$ and not in $U(x^2)$ because $U(x^2)$ contains only numbers $n \le x^2$ whose $\Omega(n)$ lies within $\varepsilon \log\log x^2$ to $\log\log x^2 = \log\log x + \log 2$. Thus the number of distinct products ab with both a and b in $U(x)$ is bounded by

$$|V(x^2)| \le \varepsilon x^2.$$

Altogether we have at most $(3\varepsilon + \varepsilon^2)x^2$ distinct products in the table. Taking ε arbitrarily small, we obtain the result. □

In 1960, P. Erdős proved that, more precisely,

$$A(x) = \frac{x^2}{(\log x)^{\alpha + o(1)}}$$

where

$$\alpha = 1 - \frac{\log(\mathrm{e} \log 2)}{\log 2} = 0.08607\ldots .$$

4.4 Remarks

Exposition in section 4.2 follows Hlawka, Schoißengeier and Taschner [10]. In section 4.3 we follow the classics of Hardy and Wright [9] and Tenenbaum [20].

Chapter 5

Congruences

After introducing basic properties of quadratic residues, we prove the quadratic reciprocity law that relates solvability of the congruences $x^2 \equiv p \pmod q$ and $x^2 \equiv q \pmod p$ for prime moduli p and q.

5.1 Quadratic residues and the quadratic reciprocity law

Let $p > 2$ be an odd prime and $a \in \mathbf{Z}$. If the congruence $x^2 \equiv a \pmod p$ has a solution $x \in \mathbf{Z}$, we say that a is a *quadratic residue* (modulo p). If it does not have solution, a is a *quadratic nonresidue* (modulo p). For example, among the nonzero residue classes $1, 2, \ldots, 10$ modulo 11, quadratic residues are $1, 3, 4, 5$, and 9 because $1^2 \equiv 1, 2^2 \equiv 4, 3^2 \equiv 9, 4^2 \equiv 5$, and $5^2 \equiv 3$ ($6^2, 7^2, \ldots$ give nothing new because $6^2 \equiv (-5)^2 \equiv 5^2$ and so on), and $2, 6, 7, 8$, and 10 are quadratic nonresidues.

Proposition. *For every prime $p > 2$ the set of nonzero residues $\mathbf{Z}_p = \{1, 2, \ldots, p-1\}$ contains the same number, $(p-1)/2$, of quadratic residues as quadratic nonresidues.*

Proof. Consider the mapping $x \mapsto x^2 \pmod p$ from $\mathbf{Z}_p$ to itself. Every $y \in \mathbf{Z}_p$ has either no preimage or exactly two because $x_1^2 \equiv x_2^2$ modulo p is equivalent with $(x_1 - x_2)(x_1 + x_2) \equiv 0$ and hence (p is a prime) with

$x_1 \equiv \pm x_2$, and $1 \not\equiv -1$ $(p > 2)$. Thus there are $(p-1)/2$ quadratic residues and $(p-1) - (p-1)/2 = (p-1)/2$ quadratic nonresidues. $\square$

Legendre's symbol $\left(\frac{a}{p}\right)$, named after Adrien-Marie Legendre (1752–1833), is for a prime $p > 2$ and $a \in \mathbf{Z}$ not divisible by p defined by

$$\left(\frac{a}{p}\right) = \begin{cases} +1 \ \dots & a \text{ is a quadratic residue} \\ -1 \ \dots & a \text{ is a quadratic nonresidue.} \end{cases}$$

For a divisible by p we set $\left(\frac{a}{p}\right) = 0$. We prove basic properties of Legendre's symbol.

Proposition. *Let $p > 2$ be a prime number and a, b be two integers.*

1. *If $a \equiv b \pmod p$ then $\left(\frac{a}{p}\right) = \left(\frac{b}{p}\right)$.*
2. *(Euler's criterion) $\left(\frac{a}{p}\right) \equiv a^{(p-1)/2} \pmod p$.*
3. *$\left(\frac{ab}{p}\right) = \left(\frac{a}{p}\right) \cdot \left(\frac{b}{p}\right)$.*

Proof. 1. This is trivial.

2. We may assume that a is not divisible by p, for else the congruence holds trivially. We have $a^{p-1} \equiv 1 \pmod p$ by Fermat's little theorem, which gives

$$(a^{(p-1)/2} - 1)(a^{(p-1)/2} + 1) \equiv 0 \pmod p.$$

Thus $a^{(p-1)/2} \equiv \pm 1 \pmod p$. If a is a quadratic residue modulo p, $a \equiv c^2$, we have $a^{(p-1)/2} \equiv c^{p-1} \equiv 1$, again by Fermat's little theorem. The congruence $a^{(p-1)/2} \equiv 1$ has no other solution besides the $(p-1)/2$ quadratic residues because the solutions are in fact roots of the polynomial $x^{(p-1)/2} - 1$ over the field $\mathbf{Z}_p$ and a well-known theorem in algebra says that the number of roots of a polynomial over a field is bounded by its degree. Thus no quadratic nonresidue b is a solution of the congruence and necessarily $b^{(p-1)/2} \equiv -1$.

3. By part 2 we have, modulo p,

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

Since the values of Legendre's symbol are $-1, 0$, and 1 (and $p > 2$), we must have equality. $\square$

Let $p > 2$ be a prime. We consider two systems of representatives of nonzero residues modulo p:

$$M = \{-\tfrac{p-1}{2}, -\tfrac{p-1}{2}+1, \ldots, -1, 1, 2, \ldots, \tfrac{p-1}{2}\} \text{ and } N = \{1, 2, \ldots, p-1\}.$$

For $a \in \mathbf{Z}$ not divisible by p we define two sequences of elements from M and N, respectively, both with length $(p-1)/2$:

$$\begin{aligned} M(a) &= (m_k \in M:\ 1 \le k \le (p-1)/2, m_k \equiv ka \pmod p), \\ N(a) &= (n_k \in N:\ 1 \le k \le (p-1)/2, n_k \equiv ka \pmod p). \end{aligned}$$

We set

$$m(a) = \#(k:\ m_k < 0) \text{ and } n(a) = \#(k:\ n_k > \tfrac{p-1}{2}).$$

It holds that $m(a) = n(a)$ because if $m_k < 0$ then $n_k = m_k + p > (p-1)/2$ and if $m_k > 0$ then $n_k = m_k \le (p-1)/2$.

Changing signs of the negative terms in $M(a)$, we obtain sequence $M(a)^+$. Similarly, $N(a)'$ is obtained from $N(a)$ by replacing every term z bigger than $(p-1)/2$ by $p-z$. Both sequences have terms in the set $\{1, 2, \ldots, (p-1)/2\}$. We claim that both $M(a)^+$ and $N(a)'$ are in fact permutations of the numbers $1, 2, \ldots, (p-1)/2$. Suppose that $M(a)^+$ is not, which means that a number appears in it twice and therefore $m_k \equiv \pm m_l$ modulo p for $1 \le k \ne l \le (p-1)/2$. But then $ka \equiv \pm la$ and $(k \pm l)a \equiv 0$, which is impossible because both factors are nonzero modulo p. The same argument works for $N(a)'$.

Proposition (Gauss' lemma). *If $p > 2$ is a prime and $a \in \mathbf{Z}$ is not divisible by p, then*

$$\left(\frac{a}{p}\right) = (-1)^{m(a)} = (-1)^{n(a)}.$$

Proof. Since $m(n) = n(n)$, it suffices to prove only the first equality. By the definition of $M(a)$ and $M(a)^+$ and the property of $M(a)^+$, modulo p we have

$$\begin{aligned} ((p-1)/2)! &= \prod_{m \in M(a)^+} m \equiv (-1)^{m(a)} \prod_{k=1}^{(p-1)/2} ka \\ &= (-1)^{m(a)}((p-1)/2)! \cdot a^{(p-1)/2}. \end{aligned}$$

Since $((p-1)/2)!$ is nonzero modulo p, we can cancel it on both sides. Using Euler's criterion (part 2 of the previous Proposition), we get

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

and the equality follows. $\square$

For example, if $p = 17$ and $a = 6$ then

$$M(6) = (6, -5, 1, 7, -4, 2, 8, -3) \text{ and } N(6) = (6, 12, 1, 7, 13, 2, 8, 14).$$

Thus $m(6) = n(6) = 3$ and $\left(\frac{6}{17}\right) = (-1)^3 = -1$.

Proposition (supplements to the reciprocity law). *Let $p > 2$ be a prime. Then*

$$\left(\frac{-1}{p}\right) = (-1)^{(p-1)/2} = \begin{cases} +1 & \dots \quad p = 4n+1 \\ -1 & \dots \quad p = 4n+3 \end{cases}$$

and

$$\left(\frac{2}{p}\right) = (-1)^{(p^2-1)/8} = \begin{cases} +1 & \dots \quad p = 8n+1, 8n+7 \\ -1 & \dots \quad p = 8n+3, 8n+5. \end{cases}$$

Proof. The first supplement follows immediately from Euler's criterion. The second supplement follows from Gauss' lemma because

$$n(2) = \frac{p-1}{2} - \#(1 \le k \le (p-1)/2 : \; 2k \le (p-1)/2) = \frac{p-1}{2} - \left\lfloor \frac{p-1}{4} \right\rfloor$$

and therefore $n(2) = 4n - 2n = 2n$ if $p = 8n+1$, $n(2) = 4n+1-2n = 2n+1$ if $p = 8n+3$, $n(2) = 4n+2-(2n+1) = 2n+1$ if $p = 8n+5$, and $n(2) = 4n+3-(2n+1) = 2n+2$ if $p = 8n+7$. $\square$

The following *quadratic reciprocity law* was known already to Euler and Legendre but the first complete proof was found by the 19 years old C. F. Gauss.

Theorem (Gauss, 1796). *Let $p, q > 2$ be two distinct odd primes. Then*

$$\left(\frac{p}{q}\right) = (-1)^{(p-1)(q-1)/4} \left(\frac{q}{p}\right) = \begin{cases} +\left(\frac{q}{p}\right) & \dots \quad p = 4m+1 \quad \textit{or} \quad q = 4n+1 \\ \\ -\left(\frac{q}{p}\right) & \dots \quad p = 4m+3 \quad \textit{and} \quad q = 4n+3. \end{cases}$$

For the proof we need two more lemmas.

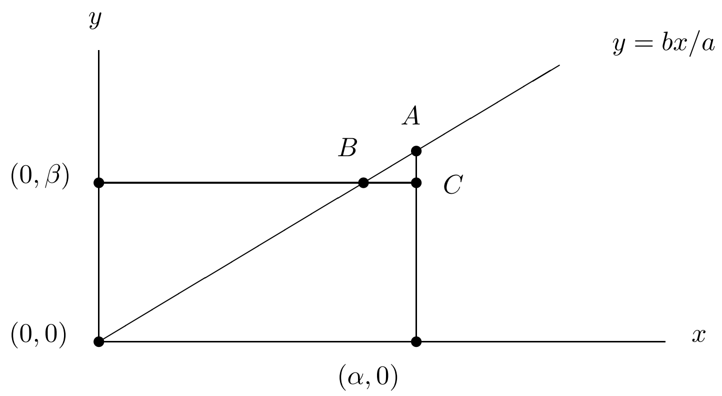
Lemma. *Let $a, b > 1$ be two distinct, odd and coprime integers. Denote*

$$S(a,b) = \sum_{i=1}^{(a-1)/2} \left\lfloor \frac{ib}{a} \right\rfloor .$$

Then

$$S(a,b) + S(b,a) = \sum_{i=1}^{(a-1)/2} \left\lfloor \frac{ib}{a} \right\rfloor + \sum_{i=1}^{(b-1)/2} \left\lfloor \frac{ia}{b} \right\rfloor = \frac{(a-1)(b-1)}{4}.$$

Proof. Let $a > b$, $\alpha = (a-1)/2$, and $\beta = (b-1)/2$. Consider this picture:



$S(a,b)$ is the number of the lattice points in the triangle with the vertices $(0,0)$, $(\alpha,0)$, $A = (\alpha, \alpha b/a)$, not counting lattice points on the x-axis. Similarly, $S(b,a)$ is the number of the lattice points in the triangle with the vertices $(0,0)$, $(0,\beta)$, $B = (\beta, \beta b/a)$, not counting lattice points on the y-axis. There are no lattice points on the segment joining $(0,0)$ and A besides $(0,0)$ (which is not counted) because a and b are coprime. Also, there are no lattice points inside the small triangle with the vertices B, A, $C = (\alpha, \beta)$ because the y-coordinate of A, $\alpha b/a = (b - b/a)/2$, lies between $(b-1)/2$

and $b/2$. Thus $S(a,b)+S(b,a)$ equals the number of the lattice points in the rectangle with the vertices $(0,0)$, $(\alpha,0)$, (α,β), $(0,\beta)$, without the lattice points on the axes, which is $(a-1)(b-1)/4$. □

Lemma. *Let $p > 2$ be a prime and $a \in \mathbf{N}$ be odd and not divisible by p. Then*

$$\left(\frac{a}{p}\right) = (-1)^{S(p,a)}$$

where $S(p,a)$ is the aforementioned sum.

Proof. Recall, for given p and a, the definition of the sequences $N(a)$ and $N(a)'$. We denote the sum of terms in $N(a)$ which are $\leq (p-1)/2$ as r and the sum of the remaining $n(a)$ terms (bigger than $(p-1)/2$) as s. Summing all terms of $N(a)'$ we get the first equation

$$\frac{p^2-1}{8} = 1 + 2 + \cdots + \frac{p-1}{2} = r + n(a)p - s$$

($N(a)'$ is a permutation of $1, 2, \ldots, (p-1)/2$). The k-th term n_k of $N(a)$ is defined by the formula

$$ka = p\left\lfloor \frac{ka}{p} \right\rfloor + n_k.$$

Summing these equations for $k = 1, 2, \ldots, (p-1)/2$, we get the second equation

$$\frac{a(p^2-1)}{8} = pS(p,a) + r + s.$$

Subtracting the first equation from the second equation we obtain

$$\frac{(a-1)(p^2-1)}{8} = p(S(p,a) - n(a)) + 2s.$$

Since a and p are odd and $(p^2-1)/8 \in \mathbf{N}$, modulo 2 this shows that $S(p,a) \equiv n(a)$. Using Gauss' lemma we conclude that $\left(\frac{a}{p}\right) = (-1)^{n(a)} = (-1)^{S(p,a)}$. □

The proof of the quadratic reciprocity law is now immediate. By the second lemma,

$$\left(\frac{q}{p}\right) = (-1)^{S(p,q)} \quad \text{and} \quad \left(\frac{p}{q}\right) = (-1)^{S(q,p)}.$$

Taking product of these equalities and using the first lemma, we get

$$\left(\frac{q}{p}\right)\left(\frac{p}{q}\right) = (-1)^{S(p,q)+S(q,p)} = (-1)^{(p-1)(q-1)/4},$$

or (because the values of the Legendre's symbol are ± 1)

$$\left(\frac{p}{q}\right) = (-1)^{(p-1)(q-1)/4}\left(\frac{q}{p}\right).$$

□

5.2 Remarks

A great number of proofs of the quadratic reciprocity law was found. We took one from Hardy and Wright [9]. For other proofs and other reciprocity laws see the book [12] by Ireland and Rosen.

Chapter 6

Integer partitions

Last chapter is devoted to the classical theory of integer partitions, which originated in the works of L. Euler in the 18th century. After introducing compositions (ordered decompositions of a number into sums of smaller numbers) and partitions (unordered decompositions), we turn our attention to partitions. We recall Ferrers diagrams and give two proofs of Euler's partition identity relating partitions with distinct parts and partitions with odd parts. Then we present a rather general identity, due to Cohen and Remmel, that subsumes Euler's identity as a very special case. This identity is proved by the inclusion-exclusion principle; in many other proofs we show the strength of generating functions. We conclude with discussion of the pentagonal identity. Its corollaries are surprising recurrences satisfied by the sequences $(p(n))_{n\geq 1}$ and $(\sigma(n))_{n\geq 1}$ where $p(n)$ counts partitions of n and $\sigma(n)$ is the sum of divisors of n.

6.1 Compositions and partitions

In how many ways can one express a natural number n as a sum of k natural numbers, regarding expressions differing by order of summands as different? In other words, what is the number of solutions $(a_1, a_2, \ldots, a_k) \in \mathbf{N}^k$ of the equation $n = a_1 + a_2 + \cdots + a_k$? The number, we denote it $c(n, k)$, is the same as the coefficient of x^n in the expansion of

$$(x + x^2 + x^3 + \cdots)^k = \left(\frac{x}{1-x}\right)^k = x^k(1-x)^{-k}.$$

By the binomial formula this equals

$$\begin{aligned} x^k \sum_{m\geq 0} \binom{-k}{m}(-x)^m &= x^k \sum_{m\geq 0} \frac{(-k)(-k-1)\dots(-k-m+1)}{m!}(-x)^m \\ &= \sum_{m\geq 0} \binom{k+m-1}{m} x^{k+m} \\ &= \sum_{r\geq k} \binom{r-1}{k-1} x^r \end{aligned}$$

and thus

$$c(n,k) = \binom{n-1}{k-1}.$$

In total, for all k we have

$$c(n) = 2^{n-1} = \sum_k \binom{n-1}{k-1}$$

expressions of n as a sum of natural summands. These expressions, in which order of summands matters, are called *compositions* of n.

Expressions of n as a sum of natural numbers, in which order of summands is irrelevant, are called *(integer) partitions of n*. We denote their number $p(n)$ and by $p(n,k)$ we denote the number of partitions of n in k parts. For example, 5 has $c(5) = 16$ compositions but only $p(5) = 7$ partitions:

$$\begin{aligned} 5 &= 5 \\ &= 4+1 \\ &= 3+2 \\ &= 3+1+1 \\ &= 2+2+1 \\ &= 2+1+1+1 \\ &= 1+1+1+1+1. \end{aligned}$$

We record partitions by weakly decreasing lists of numbers, with repetitions indicated by exponents. So the partitions of 5 are (5), $(4,1)$, $(3,2)$, $(3,1^2)$, $(2^2,1)$, $(2,1^3)$, and (1^5). The fact that λ is a partition of n is denoted as $\lambda \vdash n$.

In contrast with $c(n)$ there is no simple formula for $p(n)$. Euler noticed the identity

$$\sum_{n=0}^{\infty} p(n)x^n = \prod_{n=1}^{\infty} \frac{1}{1-x^n}$$

(we set $p(0)=1$). To see it, note that the infinite product equals

$$(x^{1\cdot 1} + x^{2\cdot 1} + x^{3\cdot 1} + \cdots)(x^{1\cdot 2} + x^{2\cdot 2} + x^{3\cdot 2} + \cdots)(x^{1\cdot 3} + x^{2\cdot 3} + x^{3\cdot 3} + \cdots)\ldots$$

and therefore, after multiplying through, the coefficient of x^n is the number of solutions $(a_1,\ldots,a_n) \in \mathbf{N}_0^n$ of the equation

$$n = a_1 + 2a_2 + 3a_3 + \cdots + na_n.$$

But this is exactly the number of ways how to express n as a sum of several (maybe none) 1's, several (maybe none) 2's, ..., several (well, one or none) n's, which by definition is $p(n)$.

More generally, for a set $A \subset \mathbf{N}$ and $n \in \mathbf{N}$ we define $p(n,A)$ to be the number of partitions of n with parts in A. Then, similarly,

$$\sum_{n=0}^{\infty} p(n,A)x^n = \prod_{n\in A} \frac{1}{1-x^n}.$$

Even more general situation is when we are given a function $f:\ \mathbf{N} \to \mathbf{N}_0$ which tells us that the number n comes in $f(n)$ colors and we define $p(n,f)$ as the number of colored partitions of n: $p(n,f)$ is the number of solutions $a_{i,j_i} \in \mathbf{N}_0$, $1 \le i \le n$, $1 \le j_i \le f(i)$ of the equation

$$n = a_{1,1} + a_{1,2} + \cdots + a_{1,f(1)} + 2(a_{2,1} + \cdots + a_{2,f(2)}) + \cdots + n(a_{n,1} + \cdots + a_{n,f(n)}).$$

Thus $p(n,A) = p(n,\chi_A)$ where χ_A is the characteristic function of the set A. Now we get

$$\sum_{n=0}^{\infty} p(n,f)x^n = \prod_{n=1}^{\infty} \frac{1}{(1-x^n)^{f(n)}}.$$

A simple but powerful tool for proving results on partitions is *Ferrers diagrams* (named after Norman MacLeod Ferrers (1829-1903)). A partition of n is visualized by a left-intended array of n dots, where the dots are grouped in rows according to the parts so that the lengths of rows weakly decrease from top to bottom. For example, the partitions $6 = 2+2+1+1$, $6 = 3+2+1$, and $17 = 5+5+5+2$ have, respectively, Ferrers diagrams

$$
\begin{array}{ll}
\bullet & \bullet \\
\bullet & \bullet \\
\bullet & \\
\bullet &
\end{array}
\qquad\qquad
\begin{array}{lll}
\bullet & \bullet & \bullet \\
\bullet & \bullet & \\
\bullet & &
\end{array}
\qquad\qquad
\begin{array}{lllll}
\bullet & \bullet & \bullet & \bullet & \bullet \\
\bullet & \bullet & \bullet & \bullet & \bullet \\
\bullet & \bullet & \bullet & \bullet & \bullet \\
\bullet & \bullet & & &
\end{array}
$$

Reading the diagram by columns instead of rows, we get the *conjugate partition* of the same number. The partitions from our example have, respectively, conjugates $6 = 4+2$, $6 = 3+2+1$ (the middle partition is self-conjugate), and $17 = 4+4+3+3+3$. Conjugation is an involutive operation: conjugate of a conjugate is the original partition. If κ and λ are conjugated partitions, then the size of the biggest part in κ is equal to the number of parts in λ. Hence conjugation is a bijection between the set of partitions of n in parts from the set $\{1, 2, \ldots, m\}$ and the set of partitions of n in at most m parts. Denoting the number of the latter partitions by $p_{\leq m}(n)$, we therefore have

$$p_{\leq m}(n) = p(n, \{1, 2, \ldots, m\}).$$

Thus

$$\sum_{n \geq 0} p_{\leq m}(n) x^n = \frac{1}{(1-x)(1-x^2)\ldots(1-x^m)}.$$

Let $p_m(n)$ be the number of partitions of n in exactly m parts. Since $p_m(n) = p_{\leq m}(n) - p_{\leq m-1}(n)$,

$$
\begin{aligned}
\sum_{n \geq 0} p_m(n) x^n &= \frac{1}{(1-x)(1-x^2)\ldots(1-x^m)} - \frac{1}{(1-x)(1-x^2)\ldots(1-x^{m-1})} \\
&= \frac{x^m}{(1-x)(1-x^2)\ldots(1-x^m)}.
\end{aligned}
$$

6.2 Euler's identity and a metaidentity of Cohen and Remmel

There is a tremendous number of identities between numbers of various kinds of partitions. We can present only few.

Theorem (Euler, 1748). *For every $n \in \mathbf{N}$, the number $r(n)$ of partitions of n in mutually distinct parts equals the number $l(n)$ of partitions of n in odd parts (which may be repeated).*

For example, 7 has 15 partitions, of which five have odd parts and five have distinct parts:

$$\begin{aligned}
7 &= 7 \text{ (odd \& distinct parts)} \\
&= 6+1 \text{ (distinct parts)} \\
&= 5+2 \text{ (distinct parts)} \\
&= 5+1+1 \text{ (odd parts)} \\
&= 4+3 \text{ (distinct parts)} \\
&= 4+2+1 \text{ (distinct parts)} \\
&= 4+1+1+1 \\
&= 3+3+1 \text{ (odd parts)} \\
&= 3+2+2 \\
&= 3+2+1+1 \\
&= 3+1+1+1+1 \text{ (odd parts)} \\
&= 2+2+2+1 \\
&= 2+2+1+1+1 \\
&= 2+1+1+1+1+1 \\
&= 1+1+1+1+1+1+1 \text{ (odd parts).}
\end{aligned}$$

We give two proofs of Euler's theorem. The third proof, by inclusion and exclusion, will be subsumed in a much more general result.

1st proof by generating functions. Since $l(n) = p(n, \{1, 3, 5, \ldots\})$, we have

$$\sum_{n \geq 0} l(n) x^n = \frac{1}{(1-x)(1-x^3)(1-x^5)\ldots}.$$

This is equal to

$$\frac{(1-x^2)(1-x^4)(1-x^6)(1-x^8)(1-x^{10})\ldots}{(1-x)(1-x^2)(1-x^3)(1-x^4)(1-x^5)(1-x^6)\ldots}$$

which simplifies to

$$(1+x)(1+x^2)(1+x^3)(1+x^4)\dots$$

because $(1-x^{2k})/(1-x^k)=1+x^k$. But it is clear that

$$\sum_{n\geq 0} r(n)x^n = (1+x)(1+x^2)(1+x^3)\dots$$

and therefore $\sum_{n\geq 0} l(n)x^n = \sum_{n\geq 0} r(n)x^n$, $l(n)=r(n)$ for every n. □

2nd proof by bijection. We construct a bijection matching the partitions of n having distinct parts with those having only odd parts. This will show that they are the same in number. Let κ be a partition of n in distinct parts and a be a part of κ. We write $a=2^b c$ where $b\in\mathbf{N}_0$ and $c\in\mathbf{N}$ is odd (this expression of a is unique) and replace a with 2^b parts c. Doing this with every part of κ, we obtain a partition λ of n with only odd parts. In the other direction, if λ is a partition of n in odd parts and a is a part of λ, we collect all a's, let λ have m of them. The number m can be expressed (in a unique way) as a sum of distinct powers of 2: $m=2^{u_1}+2^{u_2}+\cdots+2^{u_r}$ for some integers $u_1>u_2>\ldots>u_r\geq 0$ (this is the binary expansion of m). We replace the a's in λ, $a+a+\cdots+a$ (m times), by the parts $2^{u_1}a+2^{u_2}a+\cdots+2^{u_r}a$. Doing this with every part of λ, we get a partition κ of n in distinct parts (the parts are distinct because the expression "power of two $\times$ odd number" is unique). The two mappings that we defined are inverses of one another and determine the desired bijection. □

For example, for $n=7$ the described bijection matches 7 with 7, $6+1$ with $3+3+1$, $5+2$ with $5+1+1$, $4+3$ with $3+1+1+1+1$, and $4+2+1$ with $1+1+1+1+1+1+1$.

Suppose $X_1, X_2, \dots, X_k$ are finite sets, all contained in a finite superset U. Then the *principle of inclusion and exclusion (PIE)* says that

$$\begin{aligned}|U\backslash(X_1\cup\ldots\cup X_k)| &= \sum_{I\subset[k]}(-1)^{|I|}\left|\bigcap_{i\in I}X_i\right| \\ &= |U|-|X_1|-\cdots-|X_k|+|X_1\cap X_2|+\cdots\end{aligned}$$

($[k]=\{1,2,\dots,k\}$ and we define the intersection over the empty index set $I=\emptyset$ as U). From the PIE formula we immediately obtain the following identity.

Corollary. *Let $X_1, X_2, \ldots, X_k$ and $Y_1, Y_2, \ldots, Y_k$ be two k-tuples of finite subsets of a finite set U which for every $I \subset [k]$ satisfy the condition*

$$\left| \bigcap_{i \in I} X_i \right| = \left| \bigcap_{i \in I} Y_i \right|.$$

Then

$$|U \backslash (X_1 \cup \ldots \cup X_k)| = |U \backslash (Y_1 \cup \ldots \cup Y_k)|.$$

This corollary can be used to give a third proof of Euler's identity. However, this would be a far cry from using the full potential of PIE. We demonstrate it in the next general result which contains Euler's identity as a very particular case. First we need some definitions.

A *multiset* is a pair $\mathcal{A} = (A, f)$ where $A \subset \mathbf{N}$ is a finite set and $f:\ A \to \mathbf{N}$ is the *multiplicity mapping* that tells us how many times $a \in A$ appears in $\mathcal{A}$. We define its norm by

$$\|\mathcal{A}\| = \sum_{a \in A} a f(a).$$

In fact, $\mathcal{A}$ is a partition of its norm. We write

$$\mathcal{A} = (A, f) \supset \mathcal{B} = (B, g)$$

($\mathcal{A}$ *contains* $\mathcal{B}$) if $A \supset B$ and $f(a) \geq g(a)$ for every $a \in B$. If $\mathcal{A} \supset \mathcal{B}$, we define

$$\mathcal{A} - \mathcal{B} = (C, h)$$

by $C = \{a \in A :\ f(a) > g(a)\}$ and $h(a) = f(a) - g(a)$ (for $a \in A \backslash B$ we set $g(a) = 0$). For several multisets $\mathcal{A}_i = (A_i, f_i)$, $i = 1, 2, \ldots, k$, we define their union and sum by

$$\begin{aligned} \mathcal{A}_1 \cup \ldots \cup \mathcal{A}_k &= (A_1 \cup \ldots \cup A_k, \max f_i) \\ \mathcal{A}_1 + \ldots + \mathcal{A}_k &= (A_1 \cup \ldots \cup A_k, \textstyle\sum f_i) \end{aligned}$$

where $(\max f_i)(a) = \max(f_1(a), \ldots, f_k(a))$ and $(\sum f_i)(a) = f_1(a) + \cdots + f_k(a)$ (again, the undefined values $f_i(a)$ are set to 0). Observe that

$$\mathcal{A} \supset \mathcal{A}_i \text{ for every } i \in [k] \iff \mathcal{A} \supset \mathcal{A}_1 \cup \ldots \cup \mathcal{A}_k.$$

Now we can formulate a remarkable general identity, obtained by Daniel Cohen and Jeffrey Remmel.

Theorem (Cohen, 1981; Remmel, 1982). *Let $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2, \ldots)$ and $\mathcal{B} = (\mathcal{B}_1, \mathcal{B}_2, \ldots)$ be two infinite sequences of multisets (partitions) which for every finite $I \subset \mathbf{N}$ satisfy the condition*

$$\left\| \bigcup_{i \in I} \mathcal{A}_i \right\| = \left\| \bigcup_{i \in I} \mathcal{B}_i \right\| .$$

Then for every $n \in \mathbf{N}$,

$$\#\{\lambda \vdash n : \ \lambda \not\supset \mathcal{A}_i \ \textit{for}\ i = 1, 2, \ldots\} = \#\{\lambda \vdash n : \ \lambda \not\supset \mathcal{B}_i \ \textit{for}\ i = 1, 2, \ldots\},$$

that is, the number of the partitions of n containing no partition from the sequence $\mathcal{A}$ equals the number of the partitions of n containing no partition from the sequence $\mathcal{B}$.

Proof. Let $n \in \mathbf{N}$ be fixed and U be the set of all partitions λ of n, $X_i = \{\lambda \in U : \ \lambda \supset \mathcal{A}_i\}$, and $Y_i = \{\lambda \in U : \ \lambda \supset \mathcal{B}_i\}$. For a finite set of indices $1 \leq i_1 < i_2 < \ldots < i_k$ we consider the sets

$$R = X_{i_1} \cap X_{i_2} \cap \ldots \cap X_{i_k} \text{ and } S = Y_{i_1} \cap Y_{i_2} \cap \ldots \cap Y_{i_k}.$$

By the above observation,

$$R = \{\lambda \in U : \ \lambda \supset \mathcal{A}_{i_1} \cup \ldots \cup \mathcal{A}_{i_k}\} \text{ and } S = \{\lambda \in U : \ \lambda \supset \mathcal{B}_{i_1} \cup \ldots \cup \mathcal{B}_{i_k}\}.$$

For every $\lambda \in R$, the partition

$$\lambda' = (\lambda - \mathcal{A}_{i_1} \cup \ldots \cup \mathcal{A}_{i_k}) + \mathcal{B}_{i_1} \cup \ldots \cup \mathcal{B}_{i_k}$$

lies in S. Similarly, for every $\kappa \in S$, the partition

$$\kappa' = (\kappa - \mathcal{B}_{i_1} \cup \ldots \cup \mathcal{B}_{i_k}) + \mathcal{A}_{i_1} \cup \ldots \cup \mathcal{A}_{i_k}$$

lies in R. (The condition on the sequences $\mathcal{A}$ and $\mathcal{B}$ ensures that the norm of the partition (multiset) is not changed when we subtract $\mathcal{A}_i$s and then add the corresponding $\mathcal{B}_i$s). These mappings $\lambda \mapsto \lambda'$ and $\kappa \mapsto \kappa'$ are inverses of one another and they establish a bijection between R and S. Thus, for every finite set of indices $1 \leq i_1 < i_2 < \ldots < i_k$, $|R| = |S|$. By the above corollary of PIE we have $|U\backslash(X_1 \cup X_2 \cup \ldots)| = |U\backslash(Y_1 \cup Y_2 \cup \ldots)|$ (both unions are effectively finite) which we wanted to prove. $\square$

There is a simple method to construct nontrivial pairs of sequences $\mathcal{A}$ and $\mathcal{B}$ satisfying the condition in the theorem. Note that for mutually disjoint multisets $\mathcal{A}_1, \ldots, \mathcal{A}_k$ (this means that $A_i \cap A_j = \emptyset$ for $i \neq j$) the union is equal to the sum and the norm of the union is the sum of the norms. So if $\mathcal{A}$ and $\mathcal{B}$ are sequences of pairwise disjoint multisets then the condition is satisfied if and only if $\|\mathcal{A}_i\| = \|\mathcal{B}_i\|$ for every $i \in \mathbf{N}$. We call this the *disjoint satisfaction* of the condition. Thus to obtain a nontrivial pair $\mathcal{A}$ and $\mathcal{B}$ satisfying the condition, we simply take a sequence $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2, \ldots)$ of pairwise disjoint multisets and obtain $\mathcal{B}$ by splitting some of the parts in some $\mathcal{A}_i$s so that disjointness is preserved. Of course, the condition may be satisfied in some more complicated, not necessarily disjoint, ways.

We give four instances of the metaidentity. In the first three examples the condition on $\mathcal{A}$ and $\mathcal{B}$ is satisfied disjointly and in the last example it is satisfied in a more complicated way.

- (Glaisher's identity[1]) For $d \in \mathbf{N}$, $d \geq 2$, consider

$$\begin{aligned} \mathcal{A} &= (\{d\}, \{2d\}, \{3d\}, \ldots) \\ \mathcal{B} &= (\{1, 1, \ldots, 1\}, \{2, 2, \ldots, 2\}, \{3, 3, \ldots, 3\}, \ldots) \end{aligned}$$

  where in $\mathcal{B}$ all multiplicities are d. *Every n has as many partitions in parts not divisible by d as partitions in which no part appears more than $d-1$ times.* For $d = 2$ this is precisely Euler's identity.

- Consider

$$\begin{aligned} \mathcal{A} &= (\{1\}, \{4\}, \{9\}, \{16\}, \ldots) \\ \mathcal{B} &= (\{1\}, \{2, 2\}, \{3, 3, 3\}, \{4, 4, 4, 4\}, \ldots). \end{aligned}$$

  Every n has as many partitions in which no part is a square as partitions in which every part m appears at most $m-1$ times.

- (Schur's identity[2]) Consider

$$\begin{aligned} \mathcal{A} &= (\{2\}, \{3\}, \{4\}, \{6\}, \{8\}, \{9\}, \{10\}, \{12\}, \{14\}, \ldots) \\ \mathcal{B} &= (\{1, 1\}, \{3\}, \{2, 2\}, \{6\}, \{4, 4\}, \{9\}, \{5, 5\}, \{12\}, \{7, 7\}, \ldots). \end{aligned}$$

[1] Named after John Glaisher (1848-1928).
[2] Named after Issai Schur (1875-1941).

Every n has as many partitions in parts $\equiv \pm 1 \pmod 6$ as partitions in distinct parts $\equiv \pm 1 \pmod 3$.

- Consider

$$\begin{aligned} \mathcal{A} &= (\{1,1,1,1\},\{1,1,2,2\},\{2,2,2,2\},\{2,2,3,3,\},\{3,3,3,3\},\ldots) \\ \mathcal{B} &= (\{2,2\},\{2,4\},\{4,4\},\{4,6\},\{6,6\},\ldots). \end{aligned}$$

Every n has as many partitions in which parts are repeated at most thrice and in which in every two consecutive parts one of them is not repeated, as partitions in which even parts differ at least by 4 (and are not repeated, odd parts are not restricted).

6.3 The pentagonal identity

The last topic in the course is Euler's pentagonal identity. *Pentagonal numbers* are numbers

$$1, 2, 5, 7, 12, 15, 22, 26, \ldots$$

which are obtained as initial sums of two arithmetic progressions with difference 3: $5 = 1+4$, $12 = 1+4+7$, $22 = 1+4+7+10$, etc. and $7 = 2+5$, $15 = 2+5+8$, $26 = 2+5+8+11$ etc. Explicitly, pentagonal numbers are given by the formulae

$$\frac{3m^2+m}{2} \text{ for } m \in \mathbf{Z}\backslash\{0\} \text{ or } \frac{3m^2 \pm m}{2} \text{ for } m \in \mathbf{N}.$$

They count dots in diagrams of nested pentagons (which I am unable to draw) and hence their name. If one should remember only one partition identity, then it must be the following celebrated *Euler's pentagonal identity.*

Theorem (Euler, 1750). *Let $\omega(m) = (3m^2+m)/2$. The following three statements hold (and express the same fact in different formulations).*

1.

$$\prod_{n=1}^{\infty}(1-x^n) = 1 + \sum_{m=1}^{\infty}(-1)^m(x^{\omega(m)} + x^{\omega(-m)}) = \sum_{m=-\infty}^{\infty}(-1)^m x^{\omega(m)}.$$

2. *Nonpentagonal $n \in \mathbf{N}$ has as many partitions in an even number of distinct parts as partitions in an odd number of distinct parts. For pentagonal $n = \omega(\pm m)$ the number of the former partitions exceeds the number of the latter partitions by $(-1)^m$.*

3. *For every $n \in \mathbf{N}$ we have the recurrence ($p(n)$ is the number of partitions of n)*

   $$p(n) = p(n-1)+p(n-2)-p(n-5)-p(n-7)+p(n-12)+p(n-15)-\cdots$$

   where we set $p(m) = 0$ for $m < 0$ and $p(0) = 1$.

First, let us show that parts 1, 2, and 3 are saying the same thing in different words. Multiplying through the infinite product in 1, we get (as in the 1st proof of Euler's identity) that

$$\prod_{n=1}^{\infty}(1-x^n) = 1 + \sum_{n\geq 1} r^{\pm}(n)x^n$$

where $r^{\pm}(n)$ is the number of partitions of n in an even number of distinct parts minus the number of partitions of n in an odd number of distinct parts. Now it is clear that parts 1 and 2 are saying the same thing.

To see that $2 \iff 3$, recall that

$$1 + \sum_{n=1}^{\infty} p(n)x^n = \frac{1}{(1-x)(1-x^2)(1-x^3)\dots}.$$

Moving the denominator on the other side, we get

$$\left(1 + \sum_{n\geq 1} r^{\pm}(n)x^n\right)\left(1 + \sum_{n=1}^{\infty} p(n)x^n\right) = 1$$

which is the same as

$$p(n) + r^{\pm}(1)p(n-1) + r^{\pm}(2)p(n-2) + \cdots + r^{\pm}(n)p(0) = 0$$

for $n \geq 1$ (and $p(0) = 1$). If part 2 holds then we know that $r^{\pm}(n) = 0$ for $n \neq \omega(\pm m)$ and $r^{\pm}(n) = (-1)^m$ for $n = \omega(\pm m)$ and get recurrence in part

3. If part 3 holds, we restate the recurrence in the language of power series as

$$\left(1+\sum_{m=1}^{\infty}(-1)^m(x^{\omega(m)}+x^{\omega(-m)})\right)\left(1+\sum_{n=1}^{\infty}p(n)x^n\right)=1.$$

Thus

$$1+\sum_{n\geq 1}r^{\pm}(n)x^n \text{ and } 1+\sum_{m=1}^{\infty}(-1)^m(x^{\omega(m)}+x^{\omega(-m)})$$

are both multiplicative inverses of $1+\sum_{n\geq 1}p(n)x^n$ and hence must be equal as power series, coefficient by coefficient. This gives part 2.

It remains to prove one of the claims in 1, 2, or 3. Euler proved the identity in 1 by manipulating power series. A beautiful combinatorial proof of the identity in part 2 was found by Fabian Franklin (1853-1939).

Proof of part 2 (Franklin, 1881). Let U be the set of partitions λ of n with distinct parts. We shall define a pairing P of the elements in U in pairs with these properties: (i) the pairs (κ, λ) are mutually disjoint; (ii) they cover the whole U if n is not pentagonal, for $n=\omega(\pm m)$ exactly one partition is left out by P and its number of parts has the same parity as m; and (iii) the numbers of parts of the two partitions in each pair (κ, λ) have different parities. Existence of such pairing P proves 2.

We define P by means of the base and the slope of the Ferrers diagram of a partition $\lambda \in U$. The *base* b of λ is the lowest row of dots and its size $|b|$, the number of dots in it, is the smallest part of λ. The *slope* s of λ is the longest south-west going straight segment of dots that starts in the rightmost dot in the top row. Thus if m is the biggest part of λ, the size $|s|$ of the slope is the maximum $t \in \mathbf{N}$ such that $m, m-1, m-2, \ldots m-t+1$ are parts of λ.

We define two partial (i.e., not always defined) mappings $A, B:\ U \to U$. If $\lambda \in U$ and $|s|<|b|$, we obtain $A(\lambda)$ by moving the slope of λ to the bottom to form the new base. If $|s| \geq |b|$, we obtain $B(\lambda)$ by moving the base of λ to the upper right to form the new slope. For example,

$$\begin{array}{lllllll} \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \\ \bullet & \bullet & \bullet & \bullet & & & \\ \bullet & \bullet & \bullet & & & & \end{array} \quad \text{is sent by } A \text{ to} \quad \begin{array}{llllll} \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \\ \bullet & \bullet & \bullet & \bullet & & \\ \bullet & \bullet & \bullet & & & \\ \bullet & \bullet & & & & \end{array}$$

and

$$\begin{array}{lll} \bullet & \bullet & \bullet \\ \bullet & \bullet & \\ \bullet & & \end{array} \qquad \text{is sent by } B \text{ to} \qquad \begin{array}{llll} \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & & \end{array}$$

Mappings A and B are defined correctly if the base and the slope of λ are disjoint. However, there may be a problem if they intersect. If $|s| = |b| - 1$ and $s \cap b \neq \emptyset$, by applying A we get Ferrers diagram with repeated smallest part, which is not an element of U and we must leave A undefined. If $|s| < |b| - 1$ and $s \cap b \neq \emptyset$, there is no problem with A. If $|s| > |b|$ and $s \cap b \neq \emptyset$, as in our example, there is no problem with B. However, if $|s| = |b|$ and $s \cap b \neq \emptyset$, application of B does not even produce Ferrers diagram and we leave B undefined. The former bad case occurs for the partition (we denote $|b| = m + 1$) $(m + 1) + (m + 2) + \cdots + 2m = \omega(m)$ and the latter one for ($|b| = m$) $m + (m + 1) + \cdots + (2m - 1) = \omega(-m)$.

If n is not pentagonal, for every $\lambda \in U$ exactly one of the mappings A and B is defined and applying A (B) to λ, we get $\kappa \in U$ on which B (A) is defined. Moreover, $B(A(\lambda)) = \lambda$ and $A(B(\lambda)) = \lambda$. A increases and B decreases the number of parts by one. Thus the pairing P on U consisting of the pairs $(\lambda, A(\lambda))$ and $(\lambda, B(\lambda))$ (depending on which one of A and B is defined on λ) has the required properties (i)–(iii). If n is pentagonal, $n = \omega(m)$ ($n = \omega(-m)$) for $m \in \mathbf{N}$, we have the same situation with the difference that on the single partition $(m + 1) + (m + 2) + \cdots + 2m$ ($m + (m + 1) + \cdots + (2m - 1)$) neither A nor B is defined. This partition is left out by the pairing and again (i)–(iii) are satisfied. $\square$

We conclude these lecture notes with a remarkable identity for the function $\sigma(n)$ of sum of divisors of n,

$$\sigma(n) = \sum_{d|n} d,$$

discovered again by Euler. Surprisingly, $\sigma(n)$ satisfies the same recurrence as $p(n)$! Well, not completely the same since it is a rather different function.

Theorem (Euler, 1750). *For every $n \in \mathbf{N}$ we have the recurrence*

$$\sigma(n) = \sigma(n-1) + \sigma(n-2) - \sigma(n-5) - \sigma(n-7) + \sigma(n-12) + \sigma(n-15) - \cdots$$

where $\sigma(m) = 0$ for $m < 0$ and if the term $\sigma(0)$ appears in the right side (i.e., if n is pentagonal), we define it as $\sigma(0) = \sigma(n - n) = n$.

For example, we have $\sigma(15) = 1 + 3 + 5 + 15 = 24$ and

$$\begin{aligned}\sigma(15) &= \sigma(14) + \sigma(13) - \sigma(10) - \sigma(8) + \sigma(3) + \sigma(0) \\ &= 24 + 14 - 18 - 15 + 4 + 15 \\ &= 24.\end{aligned}$$

Proof. We take the logarithmic derivative of Euler's pentagonal identity in the form 1,

$$\prod_{n=1}^{\infty}(1-x^n) = 1 + \sum_{m=1}^{\infty}(-1)^m(x^{\omega(m)} + x^{\omega(-m)}),$$

more precisely we apply the operator $-x \cdot \frac{d}{dx}\log$ on both sides and get

$$\sum_{n=1}^{\infty}\frac{nx^n}{1-x^n} = \left(\sum_{m=1}^{\infty}(-1)^{m+1}(\omega(m)x^{\omega(m)} + \omega(-m)x^{\omega(-m)})\right)/D$$

where $D = 1 + \sum_{m\geq 1}(-1)^m(x^{\omega(m)} + x^{\omega(-m)})$. The power series on the left side (called Lambert series[3]) equals

$$\sum_{n=1}^{\infty}\sum_{m=1}^{\infty} nx^{mn} = \sum_{n=1}^{\infty}\sigma(n)x^n.$$

Moving the denominator D on the other side we get

$$\begin{aligned}&\left(1 + \sum_{m\geq 1}(-1)^m(x^{\omega(m)} + x^{\omega(-m)})\right)\sum_{n=1}^{\infty}\sigma(n)x^n \\ &= \sum_{m=1}^{\infty}(-1)^{m+1}(\omega(m)x^{\omega(m)} + \omega(-m)x^{\omega(-m)}).\end{aligned}$$

Comparing the coefficients of x^n on both sides we get equation

$$\sigma(n) - \sigma(n-1) - \sigma(n-2) + \sigma(n-5) + \sigma(n-7) - \cdots = 0$$

for $n \neq \omega(\pm m)$ and

$$\sigma(n) - \sigma(n-1) - \sigma(n-2) + \sigma(n-5) + \sigma(n-7) - \cdots = (-1)^{m+1}n$$

for $n = \omega(\pm m)$. □

[3]Named after Johann Lambert (1728–1777).

6.4 Remarks

The Cohen–Remmel identity was derived in [7] and [17]. By Bell [3], Euler's pentagonal identity is mentioned first time in a letter from D. Bernoulli to Euler dated January 28, 1741 (the letters from Euler to D. Bernouli in this correspondence are not extant). It took several years before Euler could produce a proof, which he did in 1750 in a letter to Ch. Goldbach. Euler's recurrence for the sum of divisors function was mentioned by him first in 1747 in a letter to Ch. Goldbach. More information on the theory of partitions can be found in the books by Andrews [1] and by Andrews and Eriksson [2].

References

1. G. E. Andrews, *The theory of partitions*, Cambridge University Press, Cambridge, 1998 (reprint of the 1976 original).

2. G. E. Andrews and K. Eriksson, *Integer partitions*, Cambridge University Press, Cambridge, 2004.

3. J. Bell, Euler and the pentagonal number theorem, ArXiv math.HO/0510054.

4. Yu. Bilu, Catalan's conjecture (after Mihăilescu), *Astérisque* **294** (2004) 1–26.

5. Yu. Bilu, Catalan without logarithmic forms (after Bugeaud, Hanrot and Mihăilescu), *J. Théor. Nombres Bordeaux* **17** (2005) 69–85.

6. M. Bruckheimer and A. Arcavi, Farey series and Pick's area theorem, *Math. Intelligencer* **17** (1995) 64–67.

7. D. I. A. Cohen, PIE-sums: a combinatorial tool for partition theory, *J. Combin. Theory, Ser. A* **31** (1981) 223–236.

8. M. Davis, Hilbert's Tenth Problem is unsolvable, *Amer. Math. Monthly* **80** (1973) 233–269.

9. G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers. Fifth edition*, Oxford University Press, Oxford 1979.

10. E. Hlawka, J. Schoißengaier and R. Taschner, *Geometric and Analytic Number Theory*, Springer-Verlag, Berlin, 1991.

11. M. N. Huxley, Exponential sums and lattice points. III. *Proc. London Math. Soc. (3)* **87** (2003) 591–609.

12. K. Ireland and M. Rosen, *A classical introduction to modern number theory. Second edition. Graduate Texts in Mathematics, 84*, Springer-Verlag, New York, 1990.

13. J. P. Jones and Yu. V. Matijasevič, Proof of recursive unsolvability of Hilbert's Tenth Problem, *Amer. Math. Monthly* **98** (1991) 689–709.

14. S. Lang, *Algebra. Revised Third edition*, Springer-Verlag, New York, 2002.

15. Yu. V. Matiyasevich, *Hilbert's Tenth Problem*, The MIT Press, Cambridge, MA, 1993.

16. P. Mihăilescu, Primary cyclotomic units and a proof of Catalan's conjecture, *J. Reine Angew. Math.* **572** (2004) 167–195.

17. J. B. Remmel, Bijective proofs of some classical partition identities, *J. Combin. Theory, Ser. A* **33** (1982) 273–286.

18. W. M. Schmidt, *Diophantine approximation. Lecture Notes in Mathematics, 785*, Springer, Berlin, 1980.

19. R. Taylor and A. Wiles, Ring-theoretic properties of certain Hecke algebras, *Ann. of Math. (2)* **141** (1995) 553–572.

20. G. Tenenbaum, *Introduction to Analytic and Probabilistic Number Theory*, Cambridge University Press, Cambridge, U.K. 1995.

21. A. Wiles, Modular elliptic curves and Fermat's last theorem, *Ann. of Math. (2)* **141** (1995) 443–551.

22. The abc conjecture home page,
http://www.math.unicaen.fr/˜nitaj/abc.html.

Index of names

Abel, Niels, 51, 53

Bernoulli, Daniel, 82

Catalan, Eugène, 19, 20, 31
Cauchy, Augustin, 4, 5, 34
Chebyshev, Pafnuty L., 44, 48–50, 52, 55
Cohen, Daniel, 68, 75, 82

de La Valée Poussin, Charles, 50
Diophantus of Alexandria, 1, 2, 4, 14, 19–21, 24, 26, 27, 36, 84
Dirichlet, Peter, 1–7, 10, 22, 41, 42, 58
Dyson, Freeman, 14

Erdős, Pál, 47, 48, 59, 60
Euclid, 11, 45
Euler, Leonhard, 1, 3, 11, 15, 38–41, 46, 54, 62, 64, 68, 70, 72, 74, 76–82

Farey, John, 1, 4–7, 18, 32, 34, 83
Fermat, Pierre de, 3, 19, 20, 26–28, 30, 31, 46, 62, 84
Ferrers, Norman, 68, 70, 71, 79, 80
Fields, John, 14
Franklin, Fabian, 79

Gauss, Carl, 25, 37, 40, 46, 63, 64, 66
Gelfond, Alexandr O., 14
Glaisher, John, 76
Goldbach, Christian, 45, 82

Hadamard, Jacques, 50
Hardy, Godfrey, 43, 44, 56, 58, 59
Hermite, Charles, 15
Hilbert, David, 1, 15, 17, 19, 20, 83, 84
Hurwitz, Adolf, 1, 4–6, 12, 18
Huxley, Martin, 43

Jacobi, Carl, 39, 40

Lagrange, Joseph-Louis, 11, 13, 22, 23, 32, 36, 38, 40
Lambert, Johann, 81
Legendre, Adrien-Marie, 62, 64, 67
Lindemann, Ferdinand von, 17
Liouville, Joseph, 1, 12–14

Mascheroni, Lorenzo, 41, 54
Mason, R. C., 19, 28, 29
Masser, David, 30
Matiyasevich, Yuri V., 20, 31
Matthew (the apostle), 46
Mertens, Franz, 44, 52–55, 57
Mihăilescu, Preda, 21, 31, 83

Minkowski, Hermann, 32, 35–38

Oesterlé, Joseph, 30

Pell, John, 19, 21, 22, 24, 26, 31
Pythagoras, 26, 27

Ramanujan, Srinivasa, 44, 56, 58, 59
Remmel, Jeffrey, 68, 75, 82
Roth, Klaus, 14, 15, 18

Schur, Issai, 76
Siegel, Carl-Ludwig, 14
Stothers, Wilson, 19, 28, 29

Taylor, Richard, 20
Thue, Axel, 14, 15, 19, 24–26
Turán, Paul, 56

von Mangoldt, Hans, 50

Wantzel, Pierre, 46
Wiles, Andrew, 20, 28, 30, 31