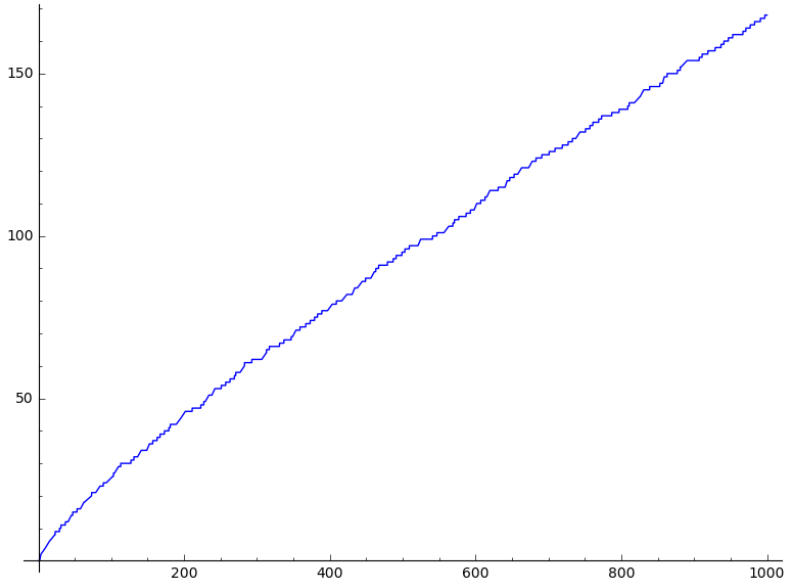
Hatice Boylan and Nils-Peter Skoruppa

Elementary Number Theory



Lecture Notes
İstanbul Üniversitesi
and Universität Siegen

Version: May 2018

Contents

Foreword

The following speech[1], given several decades ago at the occasion of the opening of one of the most famous research institutes of the world for mathematical sciences might help the reader to come closer to answer the questions *'What is number theory?'* and *'What is it good for?'*:

Why do you study number theory?

> Mathematics and German share the same disadvantage, both are universally applicable and at the same time they are the summit of artistic creation of human kind. Why do we need Goethe if we can express our wishes clearly at the market place? And for what do we need number theory if we can solve the differential equation of the heat equation numerically? Strangely enough, in this competition those domains do better which have no imaginable commercial application. One of my colleagues at Durham University was once asked by the local TV why he studies the precise dating of Crete vases, and he answers that this would be very useful for the study of the migration of

[1]This is a free translation of a German text which appeared in one of the internal publications of the Max-Planck Gesellschaft and was in turn probably translated from the original English speech. We are grateful for any hint to the English original.

the Minoan civilization. To my surprise this was accepted with respectful appreciative murmuring.

Hence our first answer to the question 'Why do you study number theory' should possibly be 'It is indispensable for the right understanding of modular forms.' After we have now put down the objections of the trifling and superficial people, we can try to answer seriously. The serious answer is, of course: 'Why not?'. Namely, beavers build dams and cuckoo borrow nests without any intent of refunding, but only humans (as far as we know) worry about the questions which prime numbers are the sum of two squares. Since we reached a partial freedom from the urgent need of surviving, the desire for knowledge and the expression of beauty were always the ultimate goal of the human race. The purpose of technology and invention is to give us time for the further study of Bach, Gauß and Goethe, and not vice versa. But it is one of the divine compensations for our existence that the compulsive quest for knowledge almost always eventually carries practical fruits.

A.O.L. Atkin,
at the occasion of a visit to the
Max-Planck-Institut for Mathematics in Bonn,
Bonn, June 1985.

Preface

These lecture notes grew out of a first course in number theory for second year students as is was given by the second author several times at the University of Siegen and by the first one in 2015/2016 at İstanbul Üniversitesi in Istanbul.

There are many books on elementary number theory, most of them in English, and with very different goals: classical, computational, theoretical, as a supplement to Algebra or from scratch. In this sense it would be unnecessary to provide a script. However, the given courses comprised each only 24 ninety minutes lecture. Hence the challenge was to reduce the contents and at the same time keep and prove rigorously key points of elementary number theory. So it might be helpful to provide a 'shortened stream-lined' version of elementary number theory as answer to the mentioned challenge. We hope that we did not do too bad when trying to reach this goal, and we hope that these lecture notes are indeed useful not only for our students, but also for our colleagues in future years.

These notes are mainly based on notes on elementary number theory which the second author collected during the past 15 years for his usage in his courses on this subject. There are some topics or treatments of such which may not be found or at least not easily found at other places. The reader will find a whole section on the basics of projective geometry since we feel that diophantine equations cannot

be treated without a certain geometric understanding. There is a whole chapter on conic sections and a natural group law on their set of rational points. This section anticipates in an elementary and easily accessible way various ideas from the theory of elliptic curves as it may be found in more advanced monographs. The theory of Pell's equation and the theory of continued fractions is here consequently explained as part of the theory of the group $\mathrm{SL}(2,\mathbb{Z})$.

Places where we want to warn the reader from wrong conclusions are marked by the first sign on the left hand side. Similarly, we occasionally leave it to the reader to find or complete an argument or computation. We indicate this by the second sign on the left side. Hints to corrections or errors are very welcome.

April 2016 Hatice Boylan and Nils-Peter Skoruppa

Chapter 1

Basics

1. The integers

We shall use $\mathbb{Z}$ for the set of integers, $\mathbb{Z}_{\geq 0}$ for the set of natural numbers[1], and $\mathbb{Q}$, $\mathbb{R}$ for the set of rational and real numbers, respectively. Recall that $\mathbb{R}$ is the smallest field containing $\mathbb{Q}$ such that every Cauchy sequence has a limit. We shall rarely deal with real numbers. Elementary number theory concerns properties of integers and rational numbers. We shall assume that the reader is acquainted with the notion of *an integer* and their basic properties, and we shall not waste time to characterize the integers axiomatically (though this would be easily possible as we shall indicate in the section "Remarks" at the end of this chapter). However, there is one property which we mention explicitly. This is the induction axiom.

Axiom (Induction Axiom)**.** *A subset of natural numbers which contains* 0*, and which contains with every number* n *also the number* $n+1$*, equals the whole set of natural numbers.*

This axiom is often applied to prove a property or identity for all natural numbers. For example one can easily prove via the induction

[1]We avoid the often used notation $\mathbb{N}$ since it is ambiguous: in the literature many authors include the number 0 in $\mathbb{N}$ whereas many others do not.

axiom that the identity

$$\sum_{k=0}^{n} k^3 = \frac{n^2(n+1)^2}{4}$$

holds true for all natural numbers.

Another important consequence, which we shall often apply is the following.

Theorem. *Every non-empty set of natural numbers has a smallest element.*

Proof. Let A be a set of natural numbers without a smallest element. We show that A is then empty. Indeed, let B be the set of natural numbers which are not in A. We have to show that B equals the set of all natural numbers, and we do this by induction. Clearly, 0 is in B since otherwise 0 would be the smallest element of A. If, for a given n, all natural $k \leq n$ are in B, then all $k \leq n+1$ are in B too since otherwise $n+1$ would be the smallest number of A. □

2. Divisibility and prime numbers

2.1. Euclid's Fundamental Theorem.

Definition. For integers a, b, we say a *divides* b, noted by $a \mid b$, if there is an $x \in \mathbb{Z}$ such that $b = ax$.

Remark. 1. The divisibility relation "'|" defines a partial ordering of $\mathbb{Z}_{\geq 0}$, i.e. "'|" is reflexive, transitive, and $a \mid b$, $b \mid a$ implies $a = b$. 2. If $d \mid a, b$, it follows $d \mid ax + by$ for all integers x and y.

Definition. A number $p \in \mathbb{Z}_{\geq 2}$ is called *prime number* (or shortly, a *prime*), if p has no other divisors than 1 and p.

Theorem (Fundamental theorem of Euclid)**.** *Every natural number possesses a unique prime factorization.*

Remark. Be prime factorization (or simply "factorization") of a number n we mean a factorization

$$n = p_1^{n_1} \cdots p_r^{n_r}$$

with prime numbers p_j and non-negative integers n_j. We can of course assume in such a writing that the exponents n_j are strictly positive and that the sequence of the p_j is strictly increasing. That such a factorization is unique means then that r, the p_j and the exponents n_j are uniquely determined by n.

Proof of the Fundamental theorem. For the 'Existence' we use induction: Let $n > 1$ be a natural number. Let p be the smallest divisor > 1 of n. If $n = p$ then n is a prime and we are done. Otherwise p and n/p possess a prime decomposition (by induction hypothesis), and so n does too.

For 'uniqueness', which we prove also by induction, we use the following fact (which is called Euclid's Lemma and whose proof will be given below): If a prime divides a product of integers, then it divides at least one of these integers. Assume that the uniqueness of prime factorization is verified for all $k < n$, and assume that n possesses prime decompositions

$$n = p_1^{n_1} \cdots p_r^{n_r} = q_1^{m_1} \cdots q_s^{m_r},$$

where $n_j, m_j \geq 1$ and $p_1 < p_2 < \cdots < p_r$ and $q_1 < q_2 < \cdots < q_s$. Then p_1 divides the product on the right hand side, hence it divides by Euclid's Lemma one of the factors, i.e. q_j for some j, and then it even equals this q_j (since q_j, as prime number, possesses only as positive divisors 1 and itself). Dividing by p_1 we obtain

$$n/p_1 = p_1^{n_1-1} \cdots p_r^{n_r} = q_1^{m_1} \cdots q_j^{m_j-1} \cdots q_s^{m_r}.$$

By induction hypothesis we find $r = s$ and $p_h = q_h$, $n_h = m_h$ for all h. □

As an immediate corollary we obtain that every positive rational number z possesses a unique prime factorization

$$z = p_1^{z_1} \cdots p_r^{z_r}$$

with primes $p_1 < \cdots < p_r$, where now, however, the integers z_j can be negative. Indeed, for seeing the existence of such a decomposition write $z = m/n$ with positive integers m and n, and replace m and n by their respective prime factorization. For the uniqueness let m be the product of all p^{z_j} with $z_j > 0$ and n be the product of all p^{-z_j}

with $z_j < 0$. Then $z = m/n$. For a given second decomposition with powers $q_j^{z_j'}$ $(0 \leq j \leq s)$ define m' and n' accordingly. Then $mn' = m'n$, and replacing here m, m', ... by the corresponding products of p^{z_j}, $q^{z_j'}$ and invoking the uniqueness of the prime factorization for integers we conclude $r = s$, $p_j = q_j$ and $z_j = z_j'$ for all j.

Let m and n be two integers, and $p_1, \ldots, p_r$ be the pairwise different primes occurring in the factorization of m and n. We can then write

$$m = p_1^{m_1} \cdots p_r^{m_r}, \quad n = p_1^{n_1} \cdots p_r^{n_r},$$

where m_j, n_j are non-negative integers, possibly equal to 0. If we have $m_j \leq n_j$ for all j, then m obviously divides n, since the quotient of m/n is a product of primes, hence an integer. The inverse is also true. If m divides n, then n/m has prime factorization $p_1^{n_1 - n_1} \cdots p_r^{n_r - m_r}$, and since it is an integer the exponents $n_j - m_j$ must all be non-negative.

The *Sieve of Eratosthenes* is an algorithm which allows to compute rapidly all primes below a given natural number n. For this one notes on a sheet of paper all natural number between 2 and n, and then one crosses out all numbers which are not primes. Namely, one starts by crossing out 2 and all multiples of 2. Then one searches for the first number which is not crossed out (which here is 3), and which is therefore a prime (since otherwise it would have a prime divisor which is smaller, but then it would be crossed out). We cross out all multiples of 3 which are strictly larger than 3. Next we look for the first number after 3 which is not crossed out (which would be 5) and which is therefore a prime (by the same argument as before). We cross out all multiples which are strictly larger. We continue in this way until we reach $\sqrt{n}$. The not crossed out numbers are then all primes $\leq n$ (since every composite number[2] $\leq n$ possesses at least a prime divisor $\leq \sqrt{n}$ and hence is already crossed out).

Theorem. *There are infinitely many prime numbers (i.e. for every integer N there exists a prime which is larger than N).*

Proof. Assume there are only finitely many prime numbers. Let P be the product of all these primes and set $n = P + 1$. Then n

[2] A positive integer is called *composite* if it is not a prime.

possesses a prime divisor p (for example, the smallest divisor of n). Since n leaves rest 1 upon dividing by any prime, but p divides n, we have a contradiction. □

A mysterious function is the *distribution of primes*

$$\pi(x) := \operatorname{card}\left(\{p \text{ prime } | p \leq x\}\right).$$

A plot of the graph of $\pi(x)$ for $0 \leq x \leq 1000$ can be found on the cover. Though $\pi(x)$ seems to follow no reasonable rule if one looks from close it seems to be rather regular in the large scale.

Theorem (Prime Number Theorem, without proof)**.** *The functions $\pi(x)$ and $\frac{x}{\log(x)}$ are asymptotically equal for $x \to \infty$ (i.e. the quotient $\pi(x)\log(x)/x$ tends to 1 for $x \to \infty$).*

2.2. Euclidean Division.

Definition (Greatest Common Divisor)**.** For integers a, b, not both zero, we call $\gcd(a,b) := \max\{d \in \mathbb{Z}_{\geq 1} : d \mid a, d \mid b\}$ the *greatest common divisor of a and b.*

Theorem. *For the* gcd *of positive a and b we have the formula*

$$\gcd(a,b) := p_1^{\min\{\alpha_1,\beta_1\}} \cdots p_r^{\min\{\alpha_r,\beta_r\}},$$

where $a = p_1^{\alpha_1} \cdots p_r^{\alpha_r}$ and $b = p_1^{\beta_1} \cdots p_r^{\beta_r}$ ($\alpha_i, \beta_i \geq 0$) denote the prime factorizations a and b.

Proof. Indeed, every divisor d of a and b must be of the form $d = p_1^{\gamma_1} \cdots p_r^{\gamma_r}$ with $\gamma_j \leq \alpha_j$ and $\gamma_j \leq \beta_j$, and vice versa, every integer of this form is a common divisor of a and b. The largest such integer is obtained by choosing γ_j equal to the minimum of α_j and β_j. This proves the theorem. □

Definition (Ideal)**.** A non-empty subset I of $\mathbb{Z}$ is called *ideal*, if for all $a, b \in I$ we have $a + b \in I$ and $a - b \in I$.

Remark. Note that 0 lies in any ideal. If a lies in an ideal, then any multiple of a also lies in the ideal.

Example. $\{0\}$ and $\mathbb{Z}$ are ideals. More generally, if d is an integer, then $\mathbb{Z}d = \{dx : x \in \mathbb{Z}\}$ is an ideal; it is called the *the principal ideal generated by d*. More generally, for arbitrary integers a_i, the set

$$\mathbb{Z}a_1 + \cdots + \mathbb{Z}a_r := \{a_1x_1 + \cdots + a_rx_r : x_1, \ldots, x_r \in \mathbb{Z}\}$$

forms an ideal.

Theorem (Principal ideal theorem)**.** *Every ideal is a principal ideal, i. e. one has $I = \mathbb{Z}d$ for a suitable d.*

For the proof we use:

Theorem (Euclidian Division)**.** *For given $m, q \in \mathbb{Z}$ and $q \neq 0$ there exist unique $x, r \in \mathbb{Z}$ such that $m = qx + r$ and $0 \leq r < |q|$.*

Example. $7 = -5 \cdot (-1) + 2$

Proof. Let r be the smallest number in

$$M := \big\{m - qx : x \in \mathbb{Z}\big\} \cap \mathbb{Z}_{\geq 0}.$$

Clearly, $m = qx+r$ and $0 \leq r$. If we had $r \geq |q|$, then $m-qx-|q| \in M$. But $m - qx - |q| < m - qx$ which is a contradiction to the minimality of $m - qx$.

The uniqueness of x and r is left as an exercise. □

Proof of the principal ideal theorem. If $I = \{0\}$ then $I = \mathbb{Z} \cdot 0$. Hence we can assume that I contains non-zero numbers. Then I contains a positive number (since with a number a it contains also $\pm a$). Let a be the smallest positive integer in I. We claim $I = \mathbb{Z}a$. Clearly, $I \supseteq \mathbb{Z}a$ (since $a \in I$). Let vice versa b in I. By Euclidean division we can write $b = xa + r$ for suitable x and $0 \leq r < a$. Writing $r = b - ax$ we see that I contains also $r < a$. But a is the smallest positive integer in I, hence $r = 0$. Therefore $b = ax$, that is $b \in \mathbb{Z}a$. □

Theorem (Bézout)**.** *For every pair $a, b \in \mathbb{Z}$, not both zero, there exist $x, y \in \mathbb{Z}$ with $ax + by = \gcd(a, b)$.*

Remark. For given a, b, c the equation $ax + by = c$ is solvable in integers x and y if and only if c divides the gcd of a and b.

Bézout's theorem is an immediate consequence of

Theorem. $\mathbb{Z}a + \mathbb{Z}b = \mathbb{Z}\gcd(a, b)$.

Proof. By the principal ideal theorem we know that the ideal $\mathbb{Z}a+\mathbb{Z}b$ is principal, that is $I := \mathbb{Z}a + \mathbb{Z}b = \mathbb{Z}g$ for a suitable positive integer g. Since a and b are in I we conclude that g divides both numbers, hence the $\gcd(a, b)$. Vice versa, the $\gcd(a, b)$ divides a and b, and hence g (since g equals $ax + by$ for suitable integers x and y). It follows $g = \gcd(a, b)$. □

Theorem (Euklid's Lemma)**.** *Let p be a prime and $a, b \in \mathbb{Z}$. Then $p \mid ab$ implies $p \mid a$ or $p \mid b$.*

Proof. Assume p does not divide a. Then $\gcd(p, a) = 1$ and hence, by Bézout's Theorem, $1 = px + ay$ for suitable x and y. Multiplying by b we obtain $b = pbx + aby$. Since p divides ab we conclude $p|b$. □

Remark. Inductively we obtain from the theorem the slightly more general statement: Is p prime, $p \mid a_1 \cdots a_r$, then $p \mid a_j$ for at least one j.

Consequence. *Note that this completes the proof of the uniqueness of the prime factorization of natural numbers.*

2.3. Euclid's algorithm. The most effective algorithm for computing the gcd of given integers is provided by *Euclid's Algorithm.* The simplest variant is based on the following lemma.

Lemma. *For all integers a, b and x, one has*

$$\gcd(a, b) = \gcd(a, b + ax).$$

Proof. Indeed, if g divides the left hand side it divides a and b and hence also a and $b+ax$, and hence the right hand side. If g divides the right hand side it divides a and $a+bx$, hence a and $b = (b+ax)-ax$, hence the left hand side. So both sides have the same divisors and are positive, so they are equal. □

Example. Successive application of the lemma (and the obvious rules $\gcd(a, b) = \gcd(b, a)$ and $\gcd(a, 0) = a$) yields an effective algorithm

for calculating the gcd of two numbers.

$$\begin{aligned}\gcd(102,27) &= \gcd(21 = 102 - 27\cdot 3, 27)\\ &= \gcd(21, 6 = 27 - 21\cdot 1)\\ &= \gcd(3 = 21 - 6\cdot 3, 6)\\ &= \gcd(3, 0 = 6 - 3\cdot 2) = 3.\end{aligned}$$

This is easy to put into a program[3]

Algorithm: Computation of the gcd of two positive integers

```
def my_first_gcd( a, b):
    while b > 0:
        c = b; b = a%b; a = c
    return a
```

We can do this also using recursion:

```
def my_second_gcd( a,b):
    return a if 0 == b else my_second_gcd(
        b, a%b)
```

If we keep track of all division steps of the preceding algorithm we can obtain at the same time also solutions x, y as in Bézout's Theorem, i.e. solutions of the equation $ax + by = \gcd(a, b)$. Namely, we start at the bottom of the last calculation, which tells us that the gcd of 102 and 27 is 3, and go up replacing at each level the remainder by the linear combination of the two preceding remainders. In our example this goes as follows:

[3]For describing algorithms we use the programming language *Python*. If you want to test or experiment with the code of this script you can easily install Python, which is freely available for almost any platform. You can, for example, install it in your Android cellphone (search in the Playstore for the app *QPython*). More advanced and also useful for other courses, you might want to use *Sage*, which is Python with mathematical libraries covering almost all parts of mathematics. If you would like to run the examples with Sage in your Web-Browser you might want to open your own Sage notebook in the *SageMathCloud* at https://cloud.sagemath.com/.

Example.

$$\begin{aligned}
3 &= 21 - \underline{6} \cdot 3 \\
&= \underline{2}1 - (27 - \underline{2}1 \cdot 1) \cdot 3 \\
&= (102 - 27 \cdot 3) - (27 - (102 - 27 \cdot 3) \cdot 1) \cdot 3 \\
&= 102 \cdot [1 + 1 \cdot 3] + 27 \cdot [-3 - 3 - 3 \cdot 1 \cdot 3] \\
&= 102 \cdot 4 + 27 \cdot (-15).
\end{aligned}$$

Replacing the remainders successively by a linear combination of the two preceding remainders is a recursive procedure. Therefore it is again extremely easy to put this into an algorithm. This can be done as follows.

Algorithm: Solving $ax + by = \gcd(a, b)$

```
def my_Bezout( a, b):
    if 0 == b: return 1,0
    x,y = my_Bezout( b, a%b)
    return y, x-(a//b)*y
```

It is sometimes useful to describe this extended Euclidean algorithm using matrices. For this we record the successive Euclidean divisions as follows:

$$\begin{aligned}
a &= a_0 b + r_1 & \left[\begin{smallmatrix} a \\ b \end{smallmatrix}\right] &= \left[\begin{smallmatrix} a_0 & 1 \\ 1 & 0 \end{smallmatrix}\right] \left[\begin{smallmatrix} b \\ r_1 \end{smallmatrix}\right] \\
b &= a_1 r_1 + r_2 & \left[\begin{smallmatrix} b \\ r_1 \end{smallmatrix}\right] &= \left[\begin{smallmatrix} a_1 & 1 \\ 1 & 0 \end{smallmatrix}\right] \left[\begin{smallmatrix} r_1 \\ r_2 \end{smallmatrix}\right] \\
r_1 &= a_2 r_2 + r_3 & \left[\begin{smallmatrix} r_1 \\ r_2 \end{smallmatrix}\right] &= \left[\begin{smallmatrix} a_2 & 1 \\ 1 & 0 \end{smallmatrix}\right] \left[\begin{smallmatrix} r_2 \\ r_3 \end{smallmatrix}\right] \\
&\vdots & &\vdots \\
r_{n-1} &= a_n r_n + 0 & \left[\begin{smallmatrix} r_{n-1} \\ r_n \end{smallmatrix}\right] &= \left[\begin{smallmatrix} a_n & 1 \\ 1 & 0 \end{smallmatrix}\right] \left[\begin{smallmatrix} r_n \\ 0 \end{smallmatrix}\right],
\end{aligned}$$

where $n > r_1 > r_2 > \cdots > r_n > 0$, and where $r_n = \gcd(a, b)$. We then have

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} a_0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a_1 & 1 \\ 1 & 0 \end{bmatrix} \cdots \begin{bmatrix} a_n & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} r_n \\ 0 \end{bmatrix}.$$

The matrix on the right is of the form $\begin{bmatrix} a/r_n & u \\ b/r_n & v \end{bmatrix}$. Its determinant is $(-1)^{n+1}$ (since the determinant of a matrix of the form $\left[\begin{smallmatrix} a & 1 \\ 1 & 0 \end{smallmatrix}\right]$ is -1). In other words, we have

$$a(-1)^{n+1}u + b(-1)^n v = \gcd(a, b),$$

which provides us with solutions x, y as in Bézout's theorem.

Definition (Least Common Multiple)**.** For $a, b \in \mathbb{Z}$, not both zero, we call the number

$$\operatorname{lcm}(a, b) := \min\{d \in \mathbb{Z}_{>0} : a|d,\ b|d\}$$

the *least common multiple of a and b.*

Theorem. *For the lcm of numbers a and b one has the formula*

$$\operatorname{lcm}(a, b) = p_1^{\max\{\alpha_1,\beta_1\}} \cdots p_r^{\max\{\alpha_r,\beta_r\}},$$

where $a = p_1^{\alpha_1} \cdot \cdots \cdot p_r^{\alpha_r}$ and $b = p_1^{\beta_1} \cdot \cdots \cdot p_r^{\beta_r}$ denote the prime factorizations of a and b.

As consequence of the formulas for the gcd and the lcm in terms of prime decompositions and the formula

$$\min\{\alpha_j, \beta_j\} + \max\{\alpha_j, \beta_j\} = \alpha_j + \beta_j$$

one obtains

Theorem. $\gcd(a, b) \cdot \operatorname{lcm}(a, b) = ab$.

Definition (GCD and LCM of more than two numbers)**.** For integers $a_1, \ldots, a_r$, not both zero, one defines

$$\begin{aligned} \gcd(a_1, \ldots, a_r) &:= \max\{d \in \mathbb{Z}_{\geq 0} : d \mid a_1, \ldots, d \mid a_r\} \\ \operatorname{lcm}(a_1, \ldots, a_r) &:= \min\{d \in \mathbb{Z}_{>0} : a_1 \mid d, \ldots, a_r \mid d\} \end{aligned}$$

Theorem. *One has the formulas*

$$\begin{aligned} a_1\mathbb{Z} + \cdots + a_r\mathbb{Z} &= \gcd(a_1, \ldots, a_r)\mathbb{Z} \\ a_1\mathbb{Z} \cap \cdots \cap a_r\mathbb{Z} &= \operatorname{lcm}(a_1, \ldots, a_r)\mathbb{Z} \end{aligned}$$

Proof. The first formula we saw already above for $r = 2$. The general case follows then on induction. The second formula we leave as an exercise. □

Remark. For $r \geq 3$ we have in general

$$\gcd(a_1 \cdots a_r) \cdot \operatorname{lcm}(a_1 \cdots a_r) \neq a_1 \cdots a_r.$$

The Euclidean algorithm as explained above can also easily be extended to more than two integers. Given a list of (say, positive) integers. one searches for the smallest one, replaces all integers by the remainder upon division by the smallest one, and then repeats this step until all but one integer are zero. An implementation could look like this.

Algorithm: Simultaneous computation of the gcd of a list of positive integers

```
def my_third_gcd( v):
    v = [ a for a in v if a != 0]
    v.sort()
    return v[0] if 1 == len(v) else
        my_third_gcd( [v[0]] + [a%v[0] for
        a in v[1:]])
```

Sometimes one can improve this algorithm slightly by modifying the Euclidean Division: instead of $b = aq+r$ with $0 \leq r < |a|$ one uses the modified Euclidean Division $b = aq' + s$ with $-|a|/2 < s \leq |a|/2$.

The careful reader might ask why we prefer the Euclidean algorithm for computing the gcd over the usual one that everybody performs in his head, namely using directly our definition of the gcd via the prime factorization of the integers in question. Here, for example $127 = 2 \cdot 3 \cdot 17$ and $27 = 3^3$, whence $\gcd(102, 27) = 3$ as we see immediately. This seems to be much shorter than the example calculation using the Euclidean algorithm which we gave above.

The answer is that this factorizing and then applying the defining formula for the gcd is, of course, the preferred method — for small

numbers though! If numbers become bigger there is a deep problem arising. Namely, we cannot factor too big numbers. There are many ingenious algorithms for factoring, but none of them is capable to surely factor an integer with a few thousand decimal digits. The most naive algorithm would try to factor a positive integer n by trying to find a divisor starting with 2, 3, etc. We can stop our search once we reach $\lfloor\sqrt{n}\rfloor$ since a composite number must obviously have at least one divisor smaller or equal to $\sqrt{n}$. If the number would be, for instance, a square of a prime we would have to perform indeed $\sqrt{n}$ many divisions before the factorization succeeds. Hence if n has 1000 decimal digits it could turn out that we have to try approximately $\sqrt{10^{1000}} = 10^{500}$ divisions before success. This is an incredibly large number as one learns if one asks a physicist. He would explain that there are approximately 10^{11} stars in our milky way. Hence it is much much faster to count the stars in our milky way than to factor a number with 1000 digits using the described naive algorithm. In fact, there are faster algorithms, but they are still all exponential in the number of digits of the given candidate for factoring. The conclusion is that we cannot be certain to compute the gcd of two numbers with several thousand digits using factorization.

But Euclid's algorithm can very well compute the gcd of two such numbers. Indeed, if we apply the Euclidean algorithm to numbers $b > a > 0$ then we expect that the remainder after the first division is in the average $a/2$. We then have to apply Euclidean division to numbers of the magnitude $a, a/2$, and we expect the remainder to be in the average $a/4$. So then we have in the third division to apply Euclidean division to numbers of the magnitude $a/2, a/4$, and we expect the remainder to be around $a/8$ etc. So we expect that the Euclidean algorithm terminates after $\log_2 a$ steps. In other words the number of divisions needed to compute the gcd of a, b is in the order of the number of binary digits of a. If a has 1000 decimal digits we would need approximately $1000 \log_2 10 \approx 3,300$ Euclidean divisions, which even the slowest smart phone would be able to perform the in a few seconds. The heuristic arguments given here can indeed be turned into a rigorous upper bound for the number of necessary divisions, and it turns out that the bound is in fact in the order of the number of binary digits of a.

We shall see later that computing gcds of big integers is indeed of practical interest, for example in cryptography based communication.

3. Congruences

3.1. Computing with congruences.

Definition. Let $a, b, m \in \mathbb{Z}$. We call a *congruent to* b *modulo* m (as formula: $a \equiv b \bmod m$ or $a \equiv_m b$), if $m \mid (a - b)$.

Remark. One has $a \equiv b \bmod 0$ if and only if $a = b$. Therefore, in the following, *the module* m will often be a non-zero integer, which allows to obtain identities modulo m which would not hold as identities for integers. Moreover, obviously $a \equiv b \bmod m$ if and only if $a \equiv b \bmod -m$. Hence, we do not loose any substantial part of theory when assuming occasionally that the module is positive.

Theorem. *For given* m *the relation* $\equiv_m$ *is an equivalence relation.*

We leave it to the reader to check for the given relation the axioms of an equivalence relation, namely that the relation is reflexive, symmetric and transitive.

Definition. The set of equivalence classes of the relation "congruent modulo m" is denoted by $\mathbb{Z}/m\mathbb{Z}$.

Theorem. *Assume* $m \neq 0$. *Then* $a \equiv b \bmod m$ *if and only if* a *and* b *leave the same remainder upon Euclidean Division by* m.

Proof. Write $a = mq + r$ and $b = mq' + r'$ with integers q, q' and $0 \leq r, r' < |m|$. Then $a - b = m(q - q') + r - r'$. From this it is clear that if m divides $a - b$ then it divides $r - r'$, and vice versa. But m divides $r - r'$ if and only if $r = r'$ as follows from $|r - r'| < |m|$. □

Hence, for $m \neq 0$, every equivalence class of the equivalence relation "congruent mod m" contains exactly one integer $0 \leq r < |m|$. Therefore, we have as many equivalence classes as residues, namely $|m|$ many. Moreover, an integer is congruent mod m to a given integer a if it is of the form $a + mx$, i.e. if it is contained in the set

$$a + m\mathbb{Z} := \{a + mx : x \in \mathbb{Z}\}.$$

Vice versa every number in this set is equivalent to a mod m. In particular, we have

$$a + m\mathbb{Z} = r + m\mathbb{Z},$$

where r is the remainder (or residue) of a after division by m. The equivalence classes of the relation "congruent mod m" are usually called *residue classes modulo* m, and the class containing a given integer a is called *residue class mod* m *of* a. We summarize:

Theorem. *Assume* $m \neq 0$.

(i) *The equivalence classes in* $\mathbb{Z}/m\mathbb{Z}$ *are of the form*

$$a + m\mathbb{Z} = \{a + mx : x \in \mathbb{Z}\}.$$

Vice versa, every such set is an equivalence class modulo m.

(ii) *One has* $\mathbb{Z}/m\mathbb{Z} = \{r + m\mathbb{Z} : 0 \leq r < m\}$. *The set* $\mathbb{Z}/m\mathbb{Z}$ *is in particular finite, one has* $\operatorname{card}(\mathbb{Z}/m\mathbb{Z}) = |m|$.

"Computing with congruences" is based on the following rules.

Theorem. *Let* m *be an integer' and assume that' for given integers* a, a', b *and* b'*' one has* $a \equiv a' \bmod m$ *and* $b \equiv b' \bmod m$. *Then*

(i) $a + b \equiv a' + b' \bmod m$, *and*

(ii) $ab \equiv a'b' \bmod m$.

Proof. By assumption we know that m divides $a - a'$ and $b - b'$. For proving (i) we write

$$(a + b) - (a' + b') = (a - a') + (b - b'),$$

from which it is obvious that the left hand side is divisible by m.

For (ii) we have to be a bit more tricky, namely we write

$$ab - a'b' = (a - a')b + a'(b - b'),$$

which again makes (ii) obvious. □

Example. We give an example for how to use the preceding rules. The reader has probably seen already the following fact:

> A given positive integer n is divisible by 9 if and only if its digit sum[4]. is divisible by 9.

[4]By *digit sum of* n one means the sum of the digits of the decimal expansion of n

For example, 123456789 is divisible by 9 since $1+2+\cdots+9 = 45$ is so. For proving the given rule we note that $10 \equiv 1 \bmod 9$. By (successive) application of (ii) this gives $10 \cdot 10 \equiv 1 \bmod 9$, $10 \cdot 10 \cdot 10 \equiv 1 \bmod 9$ and so forth. If z_l, z_{l-1}, $\ldots$, z_0 are the decimal digits of n, then

$$n = z_l \cdot 10^l + z_{l-1} \cdot 10^{l-1} + \cdots + z_1 \cdot 10 + z_0.$$

By (successive) application of (i) and (ii) we can replace here modulo 9 all powers of 10 by 1, i.e.

$$n \equiv z_l + z_{l-1} + \cdots + z_1 + z_0 \bmod 9.$$

The given rule is now obvious. Note that we have actually proved more, namely that a number leaves the same remainder upon Euclidean division by 9 as its digit sum.

We encourage the reader to work out similar division rules for division by 3, 11, 2, 4, 8, 5, 25.

We saw that we can work with congruences like with identities: We can replace in congruences modulo m left hand sides of another congruence mod m by its right hand side, we can multiply or add respective side of congruences modulo m so to obtain another congruence modulo m. However, the cancellation law does not hold true in general, i.e. from a given congruence $ka \equiv kb \bmod m$ we can in general not deduce $a \equiv b \bmod m$. For example, we have $2 \equiv -2 \bmod 4$, whereas $1 \not\equiv -1 \bmod 4$. The following theorem shows that under a certain assumption we can still apply cancellation.

Theorem. *Assume* $\gcd(k, m) = 1$. *Then* $ka \equiv kb \bmod m$ *implies* $a \equiv b \bmod m$.

Proof. By assumption and Bézout's Theorem there exist integers x and y such that $kx + my = 1$. We have to show that m divides $a - b$, and we know that m divides $k(a - b)$. For this we write

$$a - b = 1 \cdot (a - b) = (kx + my)(a - b) = kx(a - b) + my(a - b).$$

Since the right hand side is divisible by m, the claim follows. □

A residue class modulo m is called *primitive* if its members are relatively prime to m. Note that we only have to check for one member

that it is relatively prime to m to be certain that this holds true for all. We note an important special case of the preceding theorem:

Corollary. *Let p be a prime number. Then $ka = kb \bmod p$ implies $a = b \bmod p$, provided $k \not\equiv 0 \bmod p$.*

This rule resembles very much that fact that we may divide both sides of an identity for, say, rational or real numbers by a nonzero number. The deeper reason for this is revealed in algebra, where one learns that the set $\mathbb{Z}/p\mathbb{Z}$ of residue classes modulo a prime p form a *field*.

The preceding theorem can also be obtained as a consequence of the following stronger statement.

Theorem. *For given integers m and k there exists an integer k' such that $kk' \equiv 1 \bmod m$ if and only if $\gcd(k, m) = 1$.*

Proof. Assume $kk' \equiv 1 \bmod m$ for some integer k'. This means that $kk' = 1 + my$ for suitable k' and y, which implies that k and m are relatively prime. Vice versa, if $\gcd(k, m) = 1$ then by Bézout's Theorem $1 = kx+my$ for suitable x and y, and so, $kx \equiv 1 \bmod m$. □

Again we have as special case:

Corollary. *Let p be a prime number. Then, for every integer k which is not divisible by p, there exists an integer k' such that $kk' \equiv 1 \bmod p$.*

It is often necessary to calculate an inverse of an integer modulo a given m. For small m the easiest way is to try. For example, if we want to invert 2 modulo 5, we use that there are only 4 primitive residue classes modulo 5. Thea are represented by the possible non-zero residues modulo 5, i.e. by 1, 2, 3 and 4. Hence we multiply 2 by each of these until the result equals 1 modulo 5. We leave it to the reader to find the inverse of 2 mod 5 in this way. For larger modules m trying will not be possible. However, as we saw in the last proof computing the inverse modulo m of an integer amounts essentially to solve Bézout's equation $kx + my = 1$, which in turn is done by the extended Euclidean Algorithm. amounts essentially to

Algorithm: Computation of the inverse modulo m

```
def inv( k, m):
    x,y = my_Bezout( k,m) # see
        preceding section
    return x
```

There are obvious other rules for computing with congruences whose discovery and proof we leave to the reader (he will stumble over them once he starts to compute with congruences by himself). However, we mention the following. If $a \equiv b \bmod m$, then, for every integer k, we have $ak \equiv bk \bmod mk$. And vice versa, if $ka \equiv kb \bmod m$, k divides m and $k \neq 0$, then $a \equiv b \bmod m/k$.

3.2. The Chinese remainder theorem.

Theorem (Chinese Remainder Theorem)**.** *Let $m_1, \ldots, m_r$ be pairwise relatively prime positive integers. Let $a_1, \ldots, a_r \in \mathbb{Z}$. Then there is a solution x of the simultaneous congruences*

$$x = a_j \bmod m_j \qquad (1 \leq j \leq r).$$

Such a solution is modulo $m := m_1 \cdots m_r$ unique (i.e. if x' is another solution; then $x \equiv x' \bmod m$.

This theorem was indeed as far as one knows first written up in ancient China several thousand years ago. This is not surprising since the theorem, and in particular its proof, is of quite practical interest. Think of periodically recurring events (like star or planet constellations) which occur every m_1, m_2, ...years, respectively. If the first event occurred in year a_1, the second in year a_2, ..., is then there a year where all occur at the same time? The Chinese Remainder Theorem gives an affirmative answer if the periods are pairwise relatively prime. And what is the closest year in the future when all events do occur at the same time. Again, the proof of the Chinese Remainder Theorem will show how to compute this year.

Proof of the Chinese Remainder Theorem. For finding an integer x as in the theorem we set up a table as follows:

$$\begin{array}{ccccc} a_1 & m_1 & m/m_1 & m_1' & a_1(m/m_1)m_1' \\ a_2 & m_2 & m/m_2 & m_2' & a_2(m/m_2)m_2' \\ \vdots & \vdots & & & \vdots \\ a_r & m_r & m/m_r & m_r' & a_r(m/m_r)m_r' \end{array}$$

We let x be the sum of the entries of the last column, i.e. we set

$$x := a_1(m/m_1) \cdot m_1' + a_2(m/m_2) \cdot m_2' + \cdots + a_r(m/m_r) \cdot m_r'.$$

Here m_j' is an inverse modulo m_j of m/m_j, respectively, i.e. as solution of $m_j' \cdot m/m_j \equiv 1 \bmod m_j$. Note that such m_j' exist since m_j and m/m_j are relatively prime by assumption. We also know how to compute m_j' effectively as we learned in the last section. We leave it to the reader to verify that the so constructed x satisfies $x \equiv a_1 \bmod m_1$, $x \equiv a_2 \bmod m_2$ etc..

The uniqueness is easy to see: if x' is another solution, then $x \equiv x' \bmod m_j$ for all j. Therefore x-x' is divisible by all m_j, and since the m_j are pairwise relatively prime, we conclude that $x - x'$ must be divisible by the product of all m_j. □

Following the procedure described in the proof it is easy to let a computer find an x as in the theorem.

Algorithm: Solving simultaneous congruences

```
def prod(lst):
    """
    Return the product of the objects in
       the list lst.
    """
    pr = 1
    for x in lst:
        pr *= x
    return pr
```

```
def my_Chinese( d):
    """
    Return the smallest positive
        simultaneous solution x
    to the congruences
         x = d[n] mod n
    where n runs through the keys of the
        dictionary d.
    The keys must be pairwise relatively
        prime.
    """
    m = prod( d.keys())
    table = [( m/n, inv(m/n,n), d[n]) for n
         in d]
    x = sum( [a*b*c for a,b,c in table])
    return table,m,x,x%m

d = {3:2, 5:4, 7:6}
my_Chinese(d)
```

We note a theoretical consequence which is extremely important for solving congruences and for counting the solutions modulo m of a given congruence modulo m.

Let $f(x_1, \ldots, x_s)$ be a polynomial in s variables with integral coefficients, and let $m > 0$ be an integer. Assume that for each prime power $p^\alpha \parallel m$[5] we have a solution $\vec{x}_p \in \mathbb{Z}^s$ of the congruence

$$f(\vec{x}_p) \equiv 0 \bmod p^\alpha.$$

By the Chinese Remainder Theorem there exists an $\vec{x} \in \mathbb{Z}^r$, such that for every prime power $p^\alpha \parallel m$ we have $\vec{x} \equiv \vec{x}_p \bmod p^\alpha$. (These congruences are to be read and solved component by component.)

[5] We write $t \parallel m$ and call t an *exact divisor of* m, if t is a divisor of m such that $\gcd(t, m/t) = 1$.

For such an x we than have also

$$f(\vec{x}) \equiv 0 \bmod m.$$

Moreover, every solution $\vec{x}$ of the preceding congruence is obtained in such a way.

If we set

$$a(m) := \operatorname{card}\big\{(x_1, \ldots, x_r) \in \mathbb{Z} : \\ 0 \le x_1, \ldots, x_r < m, f(x_1, \ldots, x_r) \equiv 0 \bmod m\big\},$$

then the considerations of the last paragraph show

$$a(m) = \prod_{p^\alpha \| m} a(p^\alpha).$$

This formula has to be understood in the sense that the p^α run over all prime powers exactly dividing m.

It is sometimes useful to rewrite the Chinese Remainder Theorem in terms of maps. For this we define the *reduction map from* $\mathbb{Z}/m\mathbb{Z}$ *to* $\mathbb{Z}/n\mathbb{Z}$ for divisors $n|m$ as the map

$$\operatorname{red}_{m,n} : \mathbb{Z}/m\mathbb{Z} \longrightarrow \mathbb{Z}/n\mathbb{Z}, \quad a + m\mathbb{Z} \mapsto a + n\mathbb{Z}.$$

The reader should verify that this map is *well-defined*. This means the following: If we take another element b in $C := a + m\mathbb{Z}$, then $a+m\mathbb{Z} = b+m\mathbb{Z}$. Therefore, we have suddenly two definitions for $red_{m,n}(C)$, namely $a + n\mathbb{Z}$ and $b + n\mathbb{Z}$, and our definition make sense only if these two expressions define the same residue class modulo n.

Using the reduction map the Chinese Remainder Theorem can be restated as follows:

Theorem (Chinese Remainder Theorem, map theoretical formulation)**.** *Let $m_1, \ldots, m_r$ be pairwise relatively prime positive integers, and set $m = m_1 \cdots m_r$. The map*

$$\operatorname{red}_{m,m_1} \times \cdots \times \operatorname{red}_{m,m_r} : \mathbb{Z}/m\mathbb{Z} \longrightarrow \mathbb{Z}/m_1\mathbb{Z} \times \cdots \times \mathbb{Z}/m_r\mathbb{Z} \\ a + m\mathbb{Z} \mapsto (a + m_1\mathbb{Z}, \ldots, a + m_r\mathbb{Z}).$$

is bijective.

Indeed, the first part of the classical formulation of the Chinese Remainder Theorem says that our map is surjective, whereas the second part says it is injective.

3.3. Algebraic congruences mod p^n. As we saw in the discussion succeeding the Chinese remainder theorem in Section 3.2, given a polynomial $f(x_1, \dots, x_s)$ in s variables with integral coefficients, any congruence $f(x_1, \dots, x_s) \equiv 0 \bmod m$ can be reduced to the corresponding congruences modulo the prime powers p^n exactly dividing m. However, a congruence modulo a prime power p^n can often be reduced to the congruence modulo p. The advantage lies at hand. For finding a solution

$$f(x_1, \dots, x_s) \equiv 0 \bmod p$$

there is in general no better method than trying systematically all possible elements of $(\mathbb{F}_p)^s$ for being a solution. If p and s are sufficiently small such a search can be done, whereas a corresponding search modulo p^2 would already square the amount of trials. The mentioned method of reduction is an adaption of Newton's method for finding real roots of a polynomial in one variable. It is explained in the proof of the following theorem.

Theorem (Newton's method)**.** *Let f be a polynomial in s variables with integral coefficients, and p^n $(n \geq 1)$ a prime power. Assume*

$$f(x_1, \dots, x_s) \equiv 0 \bmod p^n, \quad \nabla f(x_1, \dots, x_s) \not\equiv 0 \bmod p.$$

Then there exists a solution

$$f(y_1, \dots, y_s) \equiv 0 \bmod p^{n+1} \quad \textit{with} \quad y_1, \dots, y_s \equiv x_1, \dots, x_s \bmod p^n.$$

Here ∇f is the vector of length s whose jth entry is the partial derivative of f with respect to the jth variable.

Proof of the theorem. Write $\vec{y}$ for $(y_1, \dots, y_s)$ and similar for the vector of the x_j. For the desired solution $\vec{y} \equiv \vec{x} \bmod p^n$ we make the ansatz

$$\vec{y} = \vec{x} + p^n \vec{t}$$

with a vector $\vec{t}$ so that we have to solve

$$f(\vec{x} + p^n \vec{t}) \equiv 0 \bmod p^{n+1}.$$

Here $\vec{y} = (y_1, \ldots, y_s)$ and $\vec{x} = (x_1, \ldots, x_s)$.

We expand f around $\vec{x}$ and observe that all higher terms apart from the constant and linear ones vanish modulo p^{n+1}:

$$f(\vec{x} + p^n\vec{t}) \equiv f(\vec{x}) + p^n \nabla f(\vec{x}) \cdot \vec{t} \bmod p^{n+1},$$

where the dot on the right is the usual scalar product of row vectors. The congruence of our ansatz becomes therefore

$$-\frac{1}{p^n} f(\vec{x}) \equiv \nabla f(\vec{x}) \cdot \vec{t} \bmod p.$$

But this congruence is solvable in $\vec{t}$ since we assumed that $\nabla f(\vec{x})$ is not zero modulo p. Note that the solutions $\vec{t}$ form an affine subspace of $\mathbb{F}_p^s$ of co-dimension 1. In particular, $\vec{t}$ is unique if $s = 1$. □

As we saw in the proof the case $s = 1$ is especially interesting.

Corollary. *Let f be a polynomial in one variable with integral coefficients, and p a prime number. Assume $f(y_1) \equiv 0 \bmod p$ and $f'(y_1) \not\equiv 0 \bmod p$ Then, for any n, there exists exactly one solution y_n modulo p^n of $f(y_n) \equiv 0 \bmod p^n$ with $y_n \equiv y_1 \bmod p$.*

From the uniqueness we deduce $y_{n+1} \equiv y_n \bmod p^n$. If we set $y_{n+1} = y_n + t_n p^n$, and $t_0 = y_1$, then $y_{n+1} = \sum_{\nu=0}^{n} t_\nu p^\nu$. Note that we can assume that the y_n have been chosen so that $0 \le t_\nu < p$. The sums look like the partial sums of a p-adic expansion of some object, and it is natural to ask what object this might be. The interested reader can find the answer in Section 4.2.

3.4. Primitive residue classes.

Definition. A residue class modulo m is called *primitive* if all its elements are relatively prime to m. We denote the set of primitive residue classes modulo m by $(\mathbb{Z}/m\mathbb{Z})^*$.

Remark. The reader should verify that a residue class modulo m is primitive if at least one of its elements is relatively prime to m.

Definition. *Euler's ϕ-function*[6] φ is defined on the set of positive integers and its values are

$$\varphi(m) := \operatorname{card}\left((\mathbb{Z}/m\mathbb{Z})^*\right) \qquad (m \ge 1).$$

[6] Euler's φ-function is sometimes also called *Euler's totient function*

In other words, since every residue class is represented by an integer $0 \leq r < m$, we have

$$\varphi(m) = \operatorname{card}\left(\left\{0 \leq r < m : \gcd(r, m) = 1\right\}\right),$$

or, equivalently, that $\varphi(m)$ equals the number of fractions $0 \leq x < 1$ whose denominator in shortest form is m.

Example. The first values of Euler's *phi*-function are

m	1	2	3	4	5	6	7	8	9	10	11	12	24
$\varphi(m)$	1	1	2	2	4	2	6	4	6	4	10	4	8

The table suggests the following theorem:

Theorem. *Let $m_1, \ldots, m_r$ be pairwise relatively prime positive integers, set $m = m_1 \cdots m_r$. Then $\varphi(m) = \varphi(m_1) \cdots \varphi(m_r)$.*

Proof. For this one checks that the map

$$\operatorname{red}_{m,m_1} \times \cdots \times \operatorname{red}_{m,m_r}$$

from the preceding section defines after restriction a bijection

$$(\mathbb{Z}/m\mathbb{Z})^* \longrightarrow (\mathbb{Z}/m_1\mathbb{Z})^* \times \cdots \times (\mathbb{Z}/m_r\mathbb{Z})^*.$$

From this the claimed formula is obvious. □

Lemma. *For prime powers p^α one has $\varphi(p^\alpha) = p^\alpha - p^{\alpha-1}$.*

Proof. The primitive residue classes modulo p^α are represented by those numbers from the list $0, 1, \ldots, p^\alpha - 1$ which are not divisible by p. But there are exactly $p^{\alpha-1}$ numbers in the list which are divisible by p, namely the numbers $0, p, 2p, 3p, \cdots, (p^{\alpha-1} - 1) \cdot p$. If we suppress these from the list, exactly $p^\alpha - p^{\alpha-1}$ number remain. This proves the lemma. □

The last theorem and the last lemma imply the following formula for $\varphi(m)$:

Theorem. *For any positive integer m, one has the formula*

$$\varphi(m) = m \prod_{p \mid m} \left(1 - \frac{1}{p}\right).$$

Here p runs through the (pairwise different) prime divisors of m.

Note that, for any m, we have $\varphi(m) \leq m-1$, with equality if and only if m is a prime.

After having determined the number of primitive residue classes modulo a given number m we study now a bit deeper the structure provided by these classes. The first thing to remark is that the product of two numbers which define a primitive residue class modulo m does so too. We shall tacitly apply this in the following.

Theorem. *Let m be a positive integer and a be an integer which is relatively prime to m. There exist an $n > 0$ such $a^n \equiv 1 \bmod m$.*

Proof. The powers a^k, where k runs through the positive integers cannot all be pairwise incongruent modulo m since there are at most m residue classes. Therefore there exist integers positive $k < l$ such that $a^k \equiv a^l \bmod m$. Choose an inverse a' of a modulo m, and multiply the last identity by a'^k. It follows $1 \equiv a^{l-k} \bmod m$. □

Definition. The smallest positive integer n such that $a^n \equiv 1 \bmod n$ is called the *order of a modulo m*.

Theorem. *Let m be a positive integer, a relatively prime to m and n be the order of a modulo m. Then*

$$\{a^k + m\mathbb{Z} : k \in \mathbb{Z}_{\geq 0}\} = \{a^k + m\mathbb{Z} : 0 \leq k < n\}.$$

This theorem is an immediate consequence of

Theorem. *Let a and $m > 0$ be relatively prime integers, and let n be the order of a modulo m. Then $a^k \equiv a^l \bmod m$ if and only if $k \equiv l \bmod n$.*

Proof. If $k = l + nx$ then clearly $a^k \equiv a^l \bmod m$. Assume the latter congruence. Without loss of generality we may assume $k < l$. Multiplying the congruence by a'^k for an inverse a' modulo m of a, we obtain $a^{l-k} \equiv 1 \bmod m$. Let r be the remainder of $l-k$ after division by n. Again it follows $a^r \equiv 1 \bmod m$. Since $r < n$ and n is the smallest positive integer such that $a^n \equiv 1 \bmod m$ we deduce $r = 0$, i.e. that n divides $l-k$. □

Definition. Let m be a positive integer. An integer w is called *primitive root modulo* m , if

$$\{w^n + m\mathbb{Z} : n \in \mathbb{Z}\} = (\mathbb{Z}/m\mathbb{Z})^*.$$

From the preceding theorem we deduce that a is a primitive root modulo m if and only if the order of a modulo m equals $\varphi(m)$. However, it is not at all clear whether a primitive root modulo m exists at all.

Example. The number $a := 10$ is a primitive root modulo 7: $10 \equiv_7 3$, $100 \equiv_7 2$, $1000 \equiv_7 6$, $10000 \equiv_7 4$, $100000 \equiv_7 5$, $1000000 \equiv_7 1$,

Example. The reader should check that, for every odd number a, one has $a^2 \equiv 1 \bmod 8$. Therefore the order of an odd number modulo 8 is 1 or 2. But $(\mathbb{Z}/8\mathbb{Z})^*$ has four elements. Therefore there exists no primitive root modulo 8.

We shall come back to the question which m possess primitive roots. However, for this and only because its interesting for its own sake we study, first of all, the notion of "order modulo m".

Theorem (Fermat's Little Theorem)**.** *Let p be a prime. For every integer x one has $x^p \equiv x \bmod p$.*

Proof. The claimed congruence is obviously correct if x is divisible by p. So assume that p does not divide x. We have

$$x^{p-1} \prod_{j=1}^{p-1} j = \prod_{j=1}^{p-1} (xj).$$

If we reduce both sides modulo p we observe that the product on the right hand side is congruent modulo p to the product $P := \prod_{j=1}^{p-1} j$ since the set of all numbers xj ($1 \le j \le p-1$ represents also all residue classes modulo p (indeed, if $xj \equiv xj' \bmod p$ then $j \equiv j' \bmod p$ since x is not divisible by p). It follows $x^{p-1}P \equiv P \bmod p$, and since P is not divisible by p, then $x^{p-1} \equiv 1 \bmod p$, or equivalently $x^p \equiv x \bmod p$. □

As an immediate consequence one obtains the binomial theorem modulo p.

Corollary. *For any two integers x, y, one has*

$$(x+y)^p \equiv x^p + y^p \bmod p.$$

Note that, by the usual binomial theorem, the corollary is equivalent to the statement that $\binom{p}{k}$ is divisible by p for every $1 \leq k \leq p-1$. It is not hard, however, to prove this directly by using the formula

$$\binom{p}{k} = \frac{p(p-1)\cdots(p-k+1)}{k!},$$

and noting that $k!$ is not divisible by p.

Fermat's Little Theorem should not be confused with Fermat's last Theorem, whose proof was a long outstanding problem in number theory for several hundred years and which was finally proved 20 years ago.

Theorem (Fermat's Last Theorem)**.** *The equation $a^n + b^n = c^n$ for $n > 2$ does not possess any integral solutions with $abc \neq 0$.*

Fermat's Little Theorem generalizes to arbitrary modules. Note that it implies that $x^{p-1} \equiv 1 \bmod p$ if x is not divisible by p (in fact, we used this in the proof). In this form it can be quickly generalized to arbitrary modules m.

Theorem (Euler)**.** *Let x and $m > 0$ be relatively prime integers. Then $x^{\varphi(m)} \equiv 1 \bmod m$.*

Recall that $\phi(m) = m - 1$ if m is a prime, in which case Euler's theorem becomes Fermat's Little Theorem. The proof of Euler's theorem is almost identical to the proof of Fermat's Little Theorem, and we leave the details as an exercise.

As immediate consequence we obtain:

Theorem. *Let a and $m > 0$ be relatively prime integers. Then the order of a modulo m divides $\varphi(m)$.*

Fermat's Little Theorem implies a (probabilistic) primality test: Given a positive integer m, check randomly chosen x which are relatively prime to m whether they satisfy $x^{m-1} \equiv 1 \bmod m$. If some x does not possess this test, then m cannot be a prime number. This

primality test fails however for *Carmichael numbers*. These are composite numbers m which satisfy $x^{m-1} \equiv 1 \bmod m$ for all x relatively to m. Such numbers do indeed exist. We leave it to the reader to find the first Carmichael Number.

We come back to the question which modules m possess primitive roots. We start with prime modules. Here the answer is easy.

Theorem. *Every prime p possesses a primitive root modulo p.*

For the proof we need two lemmas, which are interesting for its own sake.

Lemma. *Let p be a prime, and let $a_n, \ldots, a_0$ be integers not all divisible by p. The congruence*

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0 \equiv 0 \bmod p$$

possesses at most n solutions modulo p.

Proof. The left hand side of the congruence in question can be viewed as a polynomial, which we denote by $f(x)$. We proceed by induction over n. If $n = 0$, then $f(x) = a_0$, and since by assumption $a_0 \not\equiv 0 \bmod p$ there is no solution. Suppose now that f is a polynomial of degree $\leq n+1$ and $f(x_0) \equiv 0 \bmod p$. We shall show that f has at most n-many solutions modulo p which are different from x_0. For this, we write

$$\begin{aligned} f(x) &\equiv f(x) - f(x_0) \bmod p \\ &\equiv \sum_{k=0}^{n+1} a_k x^k - \sum_{k=0}^{n+1} a_k x_0^k \bmod p \\ &\equiv \sum_{k=0}^{n+1} a_k (x^k - x_0^k) \bmod p. \end{aligned}$$

using the formula

$$(x^k - x_0^k) = (x - x_0)(x^{k-1} + x^{k-2} x_0 + \ldots + x_0^{k-1}),$$

we obtain

$$f(x) \equiv (x - x_0)\big[a_0 + \sum_{k=1}^{n+1} a_k (x^{k-1} + x^{k-2} x_0 + \ldots + x_0^{k-1})\big] \bmod p.$$

Denote the sum on the right by $g(x)$. Then $g(x)$ defines a polynomial of degree $\leq n$. If $x_1 \not\equiv x_0 \bmod p$ and $f(x_1) \equiv 0 \bmod p$, then clearly $g(x_1) \equiv 0 \bmod p$. However, by induction hypothesis the congruence $g(x) \equiv 0 \bmod p$ possesses at most n solutions modulo p. □

Lemma. *One has*

$$\sum_{d|m} \varphi(d) = m.$$

Here the sum is over all positive divisors d of m.

Proof. Consider the m fractions

$$\frac{0}{m}, \frac{1}{m}, \frac{2}{m}, \ldots, \frac{m-1}{m}.$$

The denominator of any of these fractions in shortest form is a divisor d of m. The fractions in shortest form with denominator d are the fractions $\frac{k}{d}$, where $0 \leq k < d$ and $\gcd(k, d) = 1$. There are exactly $\varphi(d)$ many such fractions. The claimed formula is now obvious. □

Proof of the theorem on primitve roots mod primes. We need to show that there is a number w, not divisible by p, whose order modulo p equals $p - 1$. We know that the order of any a modulo p is a divisor of $p - 1$. For a divisor d of $p - 1$ let $A(d)$ denote the set of numbers $1 \leq a \leq p - 1$ whose order equals d. We shall show in a moment

$$\#A(d) \leq \varphi(d).$$

But it is clear that $\sum_{d|p-1} \#A(d) = p - 1$ (since every $1 \leq a \leq p - 1$ must occur in exactly one $A(d)$). On the other hand $\sum_{d|p-1} \varphi(d) = p-1$. Both identities are only possible if $\#A(d) = \varphi(d)$. In particular, $\#A(p - 1) = \varphi(p - 1) > 0$, which proves our theorem.

It remains to prove the claimed inequality. Assume that $A(d)$ is not empty and let a be an element of $A(d)$. We then have

$$\big\{a^k + p\mathbb{Z} : 1 \leq k \leq d\big\} = \big\{x + p\mathbb{Z} : x^d \equiv 1 \bmod p\big\}.$$

Indeed the set on the left has d elements (see theorem above), it is obviously a subset of the set on the right, and by the first lemma above the set on the right cannot have more than d elements (consider the polynomials $x^d - 1$). We conclude that $A(d)$ is contained in the left hand side, and it remains to count the powers a^k whose order modulo

p equals d. But if $a^{kt} \equiv 1 \bmod p$ if and only if d divides kt, and therefore the order of $a^k \bmod p$ equals d if and only if k is relatively prime to d. The claimed inequality follows. □

In fact, more is true:

Theorem. *For every odd prime p there exists an integer w which is a primitive root modulo any power p^α.*

Remark. The proof will actually show more. Namely, if w is a primitive root modulo p, then the order of w modulo p^2 equals $p-1$ or $p(p-1)$. In the second case w is a primitive root modulo any p^α, whereas in the first case $w+p$ is is a primitive root modulo any p^α.

Proof of the theorem. Let w be a primitive root modulo p. The order n of w modulo p^2 divides $\varphi(p^2) = p(p-1)$, and since in particular $w^n \equiv 1 \bmod p$ we see that $p-1$ divides n. Therefore $n = p(p-1)$ or $p-1$. In the latter case $(w+p)^{p-1} \equiv w^p + (p-1)w^{p-1}p \equiv 1 + (p-1)w^{p-1}p \bmod p^2$. Thus replacing w by $w+p$ if necessary we can assume that w is a primitive root modulo p^2. We claim that w is a primitive root modulo all powers of p.

For this we show by induction over α that

$$w^{\varphi(p^{\alpha-1})} = 1 + p^{\alpha-1}n_\alpha$$

with some number n_α which is not divisible by p. By choice of w this is true for $\alpha = 1$ and $\alpha = 2$. Assume it is true for some $\alpha \geq 2$. Then

$$w^{\varphi(p^\alpha)} = (1 + p^{\alpha-1}n_\alpha)^p = 1 + p^\alpha n_\alpha + \binom{p}{2} p^{2\alpha-2} n_\alpha^2 + \cdots .$$

But the third, fourth term etc. on the right is divisible by $p^{\alpha+1}$ since, for $t \geq 3$, we have $\alpha + 1 \leq t(\alpha - 1)$ (as follows from $\frac{\alpha+1}{\alpha-1} = 1 + \frac{2}{\alpha-1} \leq 3$ for $\alpha \geq 2$), and since $\alpha \leq 2\alpha - 2$ (for $\alpha \geq 2$) and $p \mid \binom{p}{2}$. (Note that the latter is not true for $p = 2$). If we write the right hand side as $1 + p^\alpha n_{\alpha+1}$ we see that $n_{\alpha+1} \equiv n_\alpha \bmod p$, i.e. that $n_{\alpha+1}$ is not divisible by p.

The claim now follows easily. Namely, let n be the order of w modulo p^α. As before we have, first of all, that n divides $\varphi(p^\alpha) =$

$p^{\alpha-1}(p-1)$, and that $p-1$ divides n. Hence $n = p^u(p-1)$ with $u \leq \alpha - 1$. Since

$$w^{p^{\alpha-2}(p-1)} = w^{\varphi(p^{\alpha-1})} = 1 + p^{\alpha-1}n_\alpha \not\equiv 1 \bmod p^\alpha$$

it follows $\alpha - 2 < u$, and therefore $u = \alpha - 1$ as was to be shown. □

The preceding theorem is not true for powers of 2 as we saw in an example above where we considered the module 8 which provides a counter example. For powers of 2 one has instead the following whose proof we leave to the ambitious reader.

Theorem. *For every odd integer a and every power 2^α there exists an $n \geq 0$ such that $a \equiv \pm 5^n \bmod 2^\alpha$.*

We finally can answer the question which positive integers m possess primitive roots.

Theorem. *A positive integer possesses a primitive root if and only if it is of the form 2, 4, p^s, or $2p^s$, where p is an odd prime and s a positive integer.*

Proof. It is obvious that 2 and 4 possess primitive roots, we proved that p^s possesses a primitive root, and any odd primitive root of p^s is one for $2p^s$.

Vice versa assume that m possesses a primitive root. Then the order of w modulo m is $\varphi(m)$. On the other hand side, by Euler's theorem $w^{\varphi(p^n)} \equiv 1 \bmod p^n$ for every prime power p^n relative prime to w. From the Chinese remainder theorem we deduce that therefore $w^N \equiv 1 \bmod m$, where N denotes the least common multiple of all $\varphi(p^n)$ ($p^n \parallel m$). Since $\varphi(m)$ is the product of all theses $\varphi(p^n)$ we conclude $N \leq \varphi(m)$, and then (since $\varphi(m)$ is the order of m) that $N = \varphi(m)$. But the latter implies that m contains at most one odd prime power (since $\varphi(p^n)$ is even for any odd p), and if m contains an odd prime power, then m cannot be divisble by 4 (since $\varphi(4) = 2$). If m is a power of 2, then it equals $m = 2$ or $m = 4$ since 8 (and accordingly any higher 2-power) possesses no primitive root. This proves the theorem. □

3.5. Sums of two squares.

Theorem (Wilson)**.** *For any prime p, one has $(p-1)! \equiv -1 \bmod p$.*

Proof. If $k \in \{1, \ldots, p-1\}$, then k is relatively prime to p, and so possesses an inverse modulo p, which after reducing modulo p is also contained in this set. We shall show in a moment that only the elements 1 and $p-1$ are their own inverses modulo p. Thus, the elements $2, \ldots, p-2$ must split up into pairs $\{x, x^{-1}\}$. It follows that their product is 1. Hence,

$$(p-1)! = 1 \cdot (p-1) \equiv -1 \bmod p.$$

It remains to prove that for $0 < k < p$, we have that $k^2 \equiv 1 \bmod p$ if and only if $k = 1$ or $k = p-1$. If $k = 1$ or $k = p-1$, then $k^2 \equiv 1 \bmod p$. Conversely, suppose that $k^2 \equiv 1 \bmod p$. Then

$$p | k^2 - 1 = (k-1)(k+1),$$

and since p is prime, $p|k-1$ or $p|k+1$. The only number in the set $\{1, \ldots, p-1\}$ which satisfies $p|k-1$ is $k = 1$, and the only number in $\{1, \ldots, p-1\}$ which satisfies $p|k+1$ is $p-1$. □

Theorem. *Let p be an odd prime. Then $x^2 \equiv -1 \bmod p$ is solvable if and only if $p \equiv 1 \bmod 4$.*

Proof. Let w denote a primitive root mod p. Recall that $-1 \equiv w^{\frac{p-1}{2}} \bmod p$. Therefore if if $\frac{p-1}{2}$ is even then $x = w^{\frac{p-1}{4}}$ is a squareroot of -1 modulo p. Vice versa, if $x^2 \equiv -1 \bmod p$ is solvable, say, with $x \equiv w^n \bmod p$, we conclude $\frac{p-1}{2} \equiv 2n \bmod p-1$, in particular, that $\frac{p-1}{2}$ must be even. □

We remark that Wilson's theorem gives us, for a prime number $p \equiv 1 \bmod 4$, a closed formula for a solution of $x^2 \equiv -1 \bmod p$, namely $x = \left(\frac{p-1}{2}\right)!$. Indeed,

$$\left(\frac{p-1}{2}\right)!^2 \equiv \left(\prod_{j=1}^{\frac{p-1}{2}} j\right)\left(\prod_{j=1}^{\frac{p-1}{2}} (p-j)\right) \equiv (p-1)! \equiv -1 \bmod p.$$

We leave it to the reader to find out where we used here that $p-1$ is divisible by 4. Note that this computation gives a second proof of the fact that $p \equiv 1 \bmod 4$ implies the solubility of the congruence

$x^2 \equiv -1 \bmod p$. Algorithmically it is, however, for big p not wise to compute a solution of $x^2 \equiv -1 \bmod p$ using this formula. If p is big this affords $(p-1)/2$ multiplications. It is better to proceed as in the first proof, namely to compute $w^{(p-1)/2}$ modulo p for some primitive root modulo p. At the first glance this seems also to afford $(p-1)/2$ multiplications. But there is a little important trick to reduce the computation of a power a^n to about $\log_2 n$ steps. This is best understood by an example: for computing a^{100} one proceeds as follows.

$$b = (a^2)^2, \quad c = ((b^2)^2)^2, \quad d = c^2, \quad a^{100} = d \cdot c \cdot b,$$

which makes 8 multiplications instead of 100. This method is sometimes called "divide and conquer". we discuss it in more detail in the section of remarks following this chapter.

Theorem (Thue)**.** *Let p be a prime. Then, for every r not divisible by p there exist numbers $0 < a, b < \sqrt{p}$ such that $b \equiv \pm ra \bmod p$. More generally, given integers $m > 0$ and $0 < A, B \leq m$, $AB > m$, then, for any r which is relatively prime to m there exist integers $0 < a < A$, $0 < b < B$ such that $b \equiv \pm ra \bmod m$.*

Proof. Consider the application which associates to each pair of integers (k, l) with $0 \leq k < A$, $0 \leq l < B$ the residue class modulo m of $kr + l$. Since there are $AB > m$ such pairs but only m residue classes modulo m, we conclude that there are two pairs $(k, l) \neq (k', l')$ such that

$$kr + l \equiv k'r + l' \bmod m.$$

Setting $a = |k - k'|$ and $b = |l - l'|$ we find $b \equiv \pm ar \bmod m$. It is clear that $|l - l'| < B$ and $|k - k'| < A$. Furthermore either $a \neq 0$ or $b \neq 0$ since (k, l) and (k', l') are different. But then also b respectively a is different from 0. Namely, if $b = 0$ then m would divide ra, and then also a (since r is relatively prime to m), which is only possible for $a = 0$. Vice versa $a = 0$ would imply that m divides b, whence $b = 0$. This proves Thue's theorem for general m. The special case for $m = p$ follows on taking $A = B = \lceil \sqrt{p} \rceil$, so that $AB > p$. □

As consequence of the two preceding theorems one obtains:

Theorem. *An odd prime p is a sum of two perfect squares if and only if $p \equiv 1 \bmod 4$.*

Proof. Indeed, assume $p = x^2 + y^2$ for two positive integers x and y. Clearly, x and y are smaller than $\sqrt{p}$. In particular, they are not divisible by p. But then we deduce from $x^2 \equiv -y^2 \bmod p$ that $(xy')^2 \equiv -1 \bmod p$, where y' is an inverse mod p of y. From the last but not least theorem we deduce $p \equiv 1 \bmod 4$.

Vice versa, if the latter is satisfied, we can solve $r^2 \equiv -1 \bmod p$. Choosing a and b for r as in Thue's Theorem we have $b^2 \equiv -a^2 \bmod p$. In other words p divides the number $n := a^2 + b^2$. But since a and b are smaller than $\sqrt{p}$ we have $n < 2p$. It follows $n = p$. □

The preceding theorem is an existence theorem, but it neither tells us how to find a decomposition of a prime as sum of two squares nor how many such decomposition's there are. It is not too hard to show that there is at most one solution $0 < x \leq y$ of $p = x^2 + y^2$. For small $p \equiv 1 \bmod 4$, for finding this solution, we can try all positive integers $x < \sqrt{p}$ until we find one such that $p - x^2$ is a perfect square. For large p this would not work since it needs to many steps. For this case we have the subsequent theorem, whose proof, however, would require methods from algebraic number theory and must therefore be skipped.

Theorem (Cornacchia)**.** *Let be a prime, $p \equiv 1 \bmod 4$, and let x a solution of $x^2 \equiv -1 \bmod p$ with $p/2 < x < p$. Denote by $\{r_n\}$ the sequence of numbers such that $r_0 = p$, $r_1 = x$ und $r_n = r_{n-2} \% r_{n-1}$[7] $(n \geq 2)$. Let l be the smallest index such that $r_l < \sqrt{p}$. Then $p - r_l^2$ is a perfect square.*

Note that an x as in the theorem always exist and can also be easily computed. Namly, choose a primitive root w mod p and compute a positive x such that $x \equiv w^{\frac{p-1}{4}} \bmod p$. Then $x^2 \equiv -1 \bmod 4$. If $x < p/2$ we replace x by $p - x$ so that then $p/2 < x < p$. As already explained above computing powers by even large exponents is no problem. This leads then to the following algorithm.

[7] For two integers a and $b \neq 0$ we use $a\%b$ for the remainder of a after division by b

Algorithm: Find a representation of a given prime p as sum of two perfect squares.

```
import math

def find_squares( p):
    """
    Return the solution (x,y) of 0<x<=y of
        p=x^2+y^2 if it exists,
    otherwise throw an exception. Input
        must be prime p = +1 mod 4.

    EXAMPLE
        >>> find_squares(7829)
        (50, 73)
        >>> find_squares(100049)
        (215, 232)
        >>> find_squares(1000037)
        (134, 991)

    """
    assert p%4 == 1, 'Error: %d must be a
        prime = 1 mod 4' % p
    # Find a solution of x^2=-1 mod p
    # using Wilson's theorem
    x = 1
    for j in range(2,(p-1)/2):
        x *= j
        x = x%p

    # Modify x if necessary so that p > x >
        p/2
    if 2*x < p:
```

```
        x = p - x

    # compute r_l as in the preceding
        theorem
    a=p; b=x
    while b*b > p:
        r=a%b; a=b; b=r

    a = int( math.sqrt(p-b*b))
    return (a,b) if a < b else (b,a)
```

4. Remarks

We end this chapter by additional material for those readers with some basic knowledge of abstract algebra.

4.1. Axiomatic characterization of the integers. In Section 1 we mentioned that it is not difficult to characterize the integers axiomatically. In fact, this can be done as follows. Let R be an *ordered ring without zero-divisors*. In other words, R is a set equipped with two binary operations "+" and "$\cdot$" which fulfill the axioms of a unitary commutative ring without zero-divisors, and there exists a subset $R_{\geq 0}$ of R which is closed under addition and multiplication, and which, for any $a \neq 0$ in R, contains either a or $-a$. That R is unitary means that there exists a multiplicative neutral element. This is unique and henceforth denoted by 1_R. One defines $a \leq b$ if $b - a$ is in $R_{\geq 0}$, and this relation defines then a *total order* on R. The integers are an example of such an ordered ring. However, there are also other examples, like for example the rational or real numbers. We assume now in addition that every subset in $R_{\geq 0}$ possesses a smallest element. We can then prove:

Theorem. *The induction axiom holds in R, i.e. $R_{\geq 0}$ is the only subset of $R_{\geq 0}$ which contains 0 and with every element a also $a + 1_R$.*

Proof. Indeed, let A be such a subset. If A was different from $R_{\geq 0}$ then $B := R_{\geq 0} \setminus A$ possesses a smallest element a_0. Clearly $a_0 > 0$. Furthermore $a_0 < 1$ (since otherwise $a_0 - 1$ would be in B). But there are no elements $0 < b < 1$ in R since for any such element b we would have $b^2 < b$ (as follows from $(1-b)b \in R_{\geq 0}$ and $(1-b)b \neq 0$ since R does not have any divisors of zero), contradicting the fact that the set of all $0 < b < 1$ would posses a minimal element if non-empty. Therefore B must be empty and $A = \mathbb{R}_{\geq 0}$. □

We can now prove that R is nothing else than the ring of integers, up to a possible different naming of its elements. More precisely, we shall prove the following theorem.

Theorem. *There exists one and only one isomorphism of rings of $\mathbb{Z}$ with R which maps $\mathbb{Z}_{\geq 0}$ onto $R_{\geq 0}$.*

Proof. Any isomorphism maps 1 to 1_R, and then any integer $n = n \cdot 1$ to $n \cdot 1_R$ (where, for negative n we mean by $n \cdot 1_R$ the element additive inverse of 1_r added $|n|$-many times to itself). Let vice versa f denote the map from $\mathbb{Z}$ to R which takes n to $n \cdot 1_R$. It is clear from the definition that this f is a homomorphism of rings. Note that 1_R is in $R_{\geq 0}$ (if -1_R was in $R_{\geq 0}$ then $1_R = (-1_R)(-1_R)$ is in $R_{\geq 0}$, a contradiction. Therefore any n-fold sum of 1_R is in $R_{\geq 0}$. In particular, f takes $\mathbb{Z}_{\geq 0}$ into $R_{\geq 0}$. The map f is injective. If $n \cdot 1_R = 0$ and $n \geq 2$, then $-1_R = (n-1) \cdot 1_r$ is in $\mathbb{R}_{\geq 0}$, a contradiction.

Finally, f is surjective since its image contains 0 and with every element a also $a + 1$, hence it contains $R_{\geq 0}$, and then consequently all of R. □

4.2. p-adic numbers. As we saw in the section on algebraic congruences mod p^n, given a polynomial $f(x)$ in one variable with integer coefficients and a number $0 \leq t_0 < p$ such that $f(t_0) \equiv 0 \bmod p$ and $f'(t_0) \not\equiv 0 \bmod p$, there exist one and only one sequence of numbers $0 \leq t_j < p$ such that $y_{n+1} := \sum_{\nu=0}^{n} t_\nu p^\nu$ satisfies $f(y_{n+1}) \equiv 0 \bmod p^{n+1}$ for all $n \geq 0$. The y_{n+1} look like the partial sums of some infinite p-adic expansion of some object, and we wondered what this object might be.

The answer to this can indeed be given. Namely, for a rational number r set $|r|_p = p^{-s}$, where s is the unique integer such that p does not occur in the factorization of r/p^s. We also set $|0|_p = 0$. The function $|\cdot|_p$ shares the same properties as the usual absolute value $|\cdot|_\infty$ on the set of rational numbers. Namely, we have $|r|_p = 0$ if and only if $r = 1$, we have $|rs|_p = |r|_p \cdot |s|_p$, and finally $|r + s|_p \leq |r|_p + |s|_p$ (in fact we even have even the stronger *ultrametric property* $|r + s|_p \leq \max(|r|_p, |s|_p)$). Using these properties one sees that the Cauchy sequences $\mathcal{C}_p$ of rational numbers with respect to the *valuation* $|\cdot|_p$ form under term-wise addition and multiplication a ring, and that the subset $\mathcal{N}_p$ of rational sequences converging to zero with respect to $|\cdot|_p$ form a ideal in $\mathcal{C}_p$. The quotient ring

$$\mathbb{Q}_p := \mathcal{C}_p/\mathcal{N}_p$$

turns out to be a field, the *field of p-adic numbers*. The map which associates to a rational number r the constant sequence with value r defines an embedding of fields (so that one identifies $\mathbb{Q}$ with its image under this embedding). The valuation $|\cdot|_p$ can be uniquely extended to all of $\mathbb{Q}_p$ so that the three properties of a valuation (and the ultrametric property) are still satisfied. The field $\mathbb{Q}_p$ is then complete with respect to $|\cdot|_p$, i.e. every Cauchy sequence of $\mathbb{Q}_p$ converges. The field $\mathbb{Q}$ is dense in $\mathbb{Q}_p$. One sets

$$\mathbb{Z}_p := \big\{x \in \mathbb{Q}_p : |x|_p \leq 1\big\}.$$

Using the ultrametric property it is easy to verify that $\mathbb{Z}_p$ is a ring, the ring of integers of $\mathbb{Q}_p$. Note that $\mathbb{Z}_p$ contains the ring $\mathbb{Z}$.

Coming back to our sequence of the y_n the congruences $y_m \equiv y_n \bmod p^n$ translate into $|y_m - y_n|_p \leq p^{-n}$, and hence our sequence is a Cauchy sequence and converges, say, towards y. In other words,

$$y = \lim_n y_{n+1} = \lim_n \sum_{\nu=0}^{n} t_\nu p^\nu,$$

and as in real analysis it is common in p-adic analysis too to denote this limit by $\sum_{\nu=0}^{\infty} t_\nu p^\nu$. Moreover, $f(y_n) \equiv 0 \bmod p^n$ translates to $|f(y_n)|_p \leq p^{-n}$, i.e. $f(y_n)$ converges to 0. Finally, it is not hard to show that polynomials are continuous functions (with respect to $|\cdot|_p$),

so that

$$f(y) = f(\lim_n y_n) = \lim_n f(y_n) = 0.$$

We can therefore state:

Theorem. *Let f be a polynomial in one variable with integral coefficients, and p a prime number. Assume $f(x) \equiv 0 \bmod p$ and $f'(x) \not\equiv 0 \bmod p$ Then there exists exactly one solution y in $\mathbb{Z}_p$ of $f(y) = 0$ with $|y - x|_p < 1$.*

Chapter 2

Higher Methods

5. Quadratic reciprocity

Given a positive integer m we want to solve quadratic congruences

$$ax^2 + bx + c \equiv 0 \bmod m.$$

We assume $a \neq 0$ so that this congruence is not linear. Multiplying this equation by $4a$ and completing the square we see that it is equivalent to

$$(2ax + b)^2 \equiv b^2 - 4ac \bmod 4am$$

The quantity

$$D := b^2 - 4ac$$

is called the discriminant of the quadratic polynomial $ax^2 + bx + c$. Recall from real analysis that the equation $ax^2 + bx + c = 0$ has no solution, one or two different solutions according as D is a non-square, zero or a non-zero square in $\mathbb{R}$, respectively. (Note that D is a non-zero square in $\mathbb{R}$ if and only if it is strictly positive). The second form of our equation shows that the role of the discriminant is similar on the level of congruences. More precisely, our problem splits up into two sub-problems: First of all, study the congruence $y^2 \equiv D \bmod 4am$ and then for any given solution y determine the set of integers modulo m such that $2ax + b \equiv y \bmod 2am$. (As an exercise the reader might verify that, for any positive integer n, the

congruence $y \equiv z \bmod 2n$ implies $y^2 \equiv z^2 \bmod 4n$.) Since we know already how to solve linear equations, i.e. how to approach the second problem, we concentrate in this section on the first one.

Definition. An integer a is called *quadratic residue modulo* m, if the congruence $x^2 \equiv a \bmod m$ is solvable. A *primitive quadratic residue modulo* m is a quadratic residue modulo m which is relatively prime to m.

As a consequence of the Chinese Remainder Theorem, it suffices to consider the case that m is a prime power p^n. For simplicity we

shall confine to odd prime powers p^n and to integers a which are not divisible by p. In this case it suffices even to study the case that m is a prime as we learn from the following theorem.

Theorem. *Let p^n be an odd prime power and a an integer which is not divisible by p. Then a is a quadratic residue modulo p^n if and only if it is a quadratic residue modulo p.*

Proof. Indeed let w be a primitive root modulo p^n. We claim that the quadratic residues modulo p^n among the powers of w are those which are even powers of w. Indeed every even power w^{2k} equals $(w^k)^2$ and is thus even a square of an integer. Vice versa if $w^k \equiv x^2 \bmod p^n$, then x is also relatively prime to p and thus $x \equiv w^l \bmod p^n$ for some integer l. It follows $w^k \equiv w^{2l} \bmod p^n$, and therefore $k \equiv 2l \bmod p^{n-1}(p-1)$, which implies that k is even.

The claim is now obvious: If a is a square modulo p then it is congruent modulo p to an even power of w (note that p is also a primitive root modulo p) and then it must also congruent modulo p^n to an even power of w. □

Remark. For $p = 2$ the preceding theorem is in general not true. The reader is encouraged to verify the following statements: Let a be odd. Then a is a quadratic residue modulo 2, it is a quadratic residue modulo 4 if and only if $a \equiv 1 \bmod 4$, and for $n \geq 3$ the number a is

a quadratic residue modulo 2^α if and only if $a \equiv 1 \bmod 8$.

We shall develop now a powerful criterion to decide whether a given integer a is a quadratic residue modulo a given odd prime p.

Definition (Legendre-Symbol)**.** For any odd prime p and any integer a set

$$\left(\frac{a}{p}\right) := \begin{cases} 1 & \text{if } x^2 \equiv a \bmod p \text{ is solvable and } \gcd(a,p) = 1, \\ 0 & \text{if } p \mid a, \\ -1 & \text{otherwise.} \end{cases}$$

Theorem. *Let p be an odd prime number.*

(1) *If $a \equiv a' \bmod p$, then we have $\left(\frac{a}{p}\right) = \left(\frac{a'}{p}\right)$.*

(2) *One has $\left(\frac{ab}{p}\right) = \left(\frac{a}{p}\right)\left(\frac{b}{p}\right)$.*

(3) $\left(\frac{1}{p}\right) = 1$ *and, for all a which are relatively prime to p, one has $\left(\frac{a^2}{p}\right) = 1$.*

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

(5) *If w is a primitive root modulo p, one has $\left(\frac{w^\nu}{p}\right) = (-1)^\nu$ for all non-negative integers ν.*

Proof. (1) and (3) are immediate from the definition of the Legendre symbol. (2) follows from (5). Indeed, we can write $a \equiv w^\nu \bmod p$ and $b \equiv w^\mu \bmod p$, so that $ab \equiv w^{\mu+\nu} \bmod p$. Therefore

$$\left(\frac{a}{p}\right)\left(\frac{b}{p}\right) = (-1)^\nu (-1)^\mu = (-1)^{\nu+\mu} = \left(\frac{ab}{p}\right).$$

(4) is also a consequence of (5). For this note that by Fermat's little theorem we have $w^{p-1} \equiv 1 \bmod p$. But then $w^{\frac{p-1}{2}}$ is a solution of $x^2 \equiv 1 \bmod p$, and we know that there are only two solutions modulo p, namely $+1$ and -1. Therefore $w^{\frac{p-1}{2}} \equiv -1 \bmod p$ or $w^{\frac{p-1}{2}} \equiv +1 \bmod p$. The latter is impossible since w is a primitive root. Hence $w^{\frac{p-1}{2}} \equiv -1 \bmod p$, and so by (5) $\left(\frac{-1}{p}\right) = (-1)^{\frac{p-1}{2}}$.

Finally, for proving (5) we have to prove that ν is even if and only if $w^\nu \equiv x^2 \bmod p$ is solvable. If ν is even then $w^\nu \equiv w^{\nu/2} \bmod p$. Suppose vice versa that $w^\nu \equiv x^2 \bmod p$ for a suitable x. We write $x \equiv w^\mu \bmod p$. It follows $w^\nu \equiv w^{2\mu} \bmod p$. Therefore $\nu \equiv 2\mu \bmod 2$, as was to be proven. □

Definition. A *Dirichlet character modulo* m is a homomorphism of groups[1] $\chi : (\mathbb{Z}/m\mathbb{Z})^* \to \mathbb{C}^*$ (i.e. a map which satisfies $\chi(ab) = \chi(a)\chi(b)$ for all $a, b \in (\mathbb{Z}/m\mathbb{Z})^*$).

Remark. Let χ be a Dirichlet character modulo m. (1) One has $\chi(1) = 1$. (2) The values of χ are always $\phi(m)$th roots of unity. In particular, a *real Dirichlet character* (i.e. a Dirichlet character which assumes only real values) takes on only the values ± 1. (3) We can associate to χ the map $\widetilde{\chi} : \mathbb{Z} \to \mathbb{C}$ which is defined as

$$\widetilde{\chi}(x) = \begin{cases} \chi(x + m\mathbb{Z}) & \text{falls } \gcd(x, m) = 1, \\ 0 & \text{sonst.} \end{cases}$$

This map has the properties that

(i) $\widetilde{\chi}(x + m) = \widetilde{\chi}(x)$,

(ii) $\widetilde{\chi}(x) = 0$ if and only if $\gcd(x, m) \neq 1$,

(iii) $\widetilde{\chi}(xy) = \widetilde{\chi}(x)\,\widetilde{\chi}(y)$ for all integers x and y.

By abuse of language, one calls, as we also shall do, a map $\widetilde{\chi} : \mathbb{Z} \to \mathbb{C}$ satisfying (i) to (iii) a Dirichlet character modulo m. This is justified by the fact that vice versa a map $\widetilde{\chi}$ satisfying (i) to (iii) gives rise to a Dirichlet character χ via $\chi(x + m\mathbb{Z}) := \widetilde{\chi}(x)$. We leave it to the reader to verify that this is well-defined and defines a Dirichlet character on the sense of the given definition.

Corollary. *The map*

$$(\mathbb{Z}/p\mathbb{Z})^* \to \{\pm 1\}, \quad a + p\mathbb{Z} \to \left(\frac{a}{p}\right)$$

is well-defined ad defines a Dirichlet character modulo p.

Another important consequence is a follows.

Corollary. *Let* w *be a primitive root modulo* p. *Then the set* $\{1 = w^0, w^2, w^4, ..., w^{p-1}\}$ *provides a complete set of representatives for the primitive quadratic residues modulo* p. *In particular there are exactly* $\frac{p-1}{2}$ *primitive quadratic residues and the same number of prmitive quadratic non-residues modulo* p.

[1] A homomorphism of groups $f : G \to H$ between groups G and H, is a map such that $f(ab) = f(a)f(b)$ for all $a, b \in G$.

Theorem (Euler's criterion)**.** *For all integers a which are not divisible by p, one has*

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

Proof. This follows from (5) of the last Theorem. Namely, let $a \equiv w^\nu \bmod p$, where w is a primitive root modulo p. Then

$$\left(\frac{a}{p}\right) = \left(\frac{w^\nu}{p}\right) = (-1)^\nu \equiv w^{\frac{p-1}{2}\nu} \equiv a^{\frac{p-1}{2}} \bmod p,$$

where we use again $w^{\frac{p-1}{2}} \equiv -1 \bmod p$ (see the proof of the last theorem). □

Theorem (Gauss' criterion)**.** *Let a be relatively prime to p. Let n denote the number of $0 < j < \frac{p}{2}$ such that $aj \equiv -j' \bmod p$ for some $0 < j' < \frac{p}{2}$. Then one has*

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

Proof. Note that the numbers j and $-j$, where $0 < j < \frac{p}{2}$, provide a system of representatives for the primitive residue classes modulo p. Moreover the map $x + p\mathbb{Z} \mapsto ax + p\mathbb{Z}$ defines a permutation of the set of primitive residue classes modulo p. Therefore, for every $0 < j < \frac{p}{2}$ there exists a $0 < j' < \frac{p}{2}$ and an ε_j in $\{\pm 1\}$ such that $aj \equiv \varepsilon_j j' \bmod p$.

After these preparations we have

$$a^{\frac{p-1}{2}} \prod_j j \equiv \prod_j aj \equiv \prod_j \varepsilon_j j' = \prod_j \varepsilon_j \prod_j j' \bmod p,$$

where the products runs over all $0 < j < \frac{p}{2}$, respectively. But $\prod_j \varepsilon_j = (-1)^n$ and $\prod_j j' = \prod_j j$, since $j \mapsto j'$ defines a permutation of the set $0 < j < \frac{p}{2}$. It follows $a^{\frac{p-1}{2}} \equiv (-1)^n$, and hence, by Euler's criterion, $\left(\frac{a}{p}\right) = (-1)^n$, which was to be proven. □

Recall that we proved the following fact:

> The number -1 is a quadratic residue modulo p if and only if p is quadratic residue modulo 4.

(Indeed, this is nothing else but a paraphrase of part (4) of the theorem immediatly following the definition of the Legendre symbol.) Gauss' criterion provides a mean to obtain statements like this for residue classes different from -1 too. Such rules are called *reciprocity laws*. We shall state the most general *reciprocity law* below. However, it is worthwhile to deduce special cases directly from Gauss' criterion.

For this, we formulate Gauss' criterion slightly differently. Assume that a is positive and not divisible by p. Fix a $0 < j < \frac{p}{2}$. Then $aj - kp$ is in the interval $(-\frac{p}{2}, 0)$ for a suitable integer k if and only if j is in $(\frac{2k-1}{2a}p, \frac{2k}{2a}p)$ for some $1 \leq k \leq \lfloor \frac{a}{2} \rfloor$. Applying Gauss' criterion we therefore obtain

Corollary. *For any positive integer a not divisible by p, one has*

$$\left(\frac{a}{p}\right) = (-1)^{\sum_{k=1}^{\lfloor \frac{a}{2} \rfloor} \#(\frac{(2k-1)p}{2a}, \frac{2kp}{2a}) \cap \mathbb{Z}}.$$

As a consequence one obtains the following reciprocity law.

Theorem. *One has* $\left(\frac{2}{p}\right) = (-1)^{\frac{p^2-1}{8}}$.

Proof. According to the formula of the last theorem we have $\left(\frac{2}{p}\right) = (-1)^N$, where N denotes the number of integers in the interval $(\frac{p}{4}, \frac{p}{2})$. We have to consider 4 cases: $p \equiv 1, 3, 5, 7 \bmod 8$, and show that $\left(\frac{2}{p}\right) = 1$ exactly in the first and last case.

Suppose that $p = 8k + 1$. Then $(\frac{p}{4}, \frac{p}{2}) = (2k + \frac{1}{4}, 4k + \frac{1}{2})$, and therefore $(\frac{p}{4}, \frac{p}{2}) \cap \mathbb{Z} = \{2k+1, 2k+2, \ldots, 4k\}$, and so indeed $N = 2k$, $\left(\frac{2}{p}\right) = +1$.

Suppose that $p = 8k + 3$. Then $(\frac{p}{4}, \frac{p}{2}) = (2k + \frac{3}{4}, 4k + \frac{3}{2})$, and therefore $(\frac{p}{4}, \frac{p}{2}) \cap \mathbb{Z} = \{2k+1, 2k+2, \ldots, 4k+1\}$, and so $N = 2k+1$, $\left(\frac{2}{p}\right) = -1$.

Similarly, if $p = 8k + 5$, then $(\frac{p}{4}, \frac{p}{2}) = (2k + \frac{5}{4}, 4k + \frac{5}{2})$, and therefore $(\frac{p}{4}, \frac{p}{2}) \cap \mathbb{Z} = \{2k + 2, 2k + 2, \ldots, 4k + 2\}$, $N = 2k + 1$, $\left(\frac{2}{p}\right) = -1$.

Finally, if $p = 8k+7$, then $(\frac{p}{4}, \frac{p}{2}) = (2k+\frac{7}{4}, 4k+\frac{7}{2})$, and therefore $(\frac{p}{4}, \frac{p}{2}) \cap \mathbb{Z} = \{2k + 2, 2k + 2, \ldots, 4k + 3\}$, $N = 2k + 2$, $\left(\frac{2}{p}\right) = +1$. □

Example. In a similar way as the preceding theorem the reader should try to prove for any prime $p \geq 5$:

$$\left(\frac{3}{p}\right) = \left(\frac{p}{3}\right)(-1)^{\frac{p-1}{2}}.$$

We finally prove the general reciprocity law for two odd prime numbers. For this we note, first of all, another consequence of Gauss' criterion.

Theorem. *Let a be an odd integer not divisible by p. Then one has*

$$\left(\frac{a}{p}\right) = (-1)^{\sum_{j=1}^{\frac{p-1}{2}} \lfloor \frac{aj}{p} \rfloor}.$$

Proof. With the notations as in the proof of Gauss' criterion we have for $0 < j < \frac{p}{2}$

$$aj = p\left\lfloor \frac{aj}{p} \right\rfloor + j' \qquad \text{if } \varepsilon_j = +1,$$

$$aj = p\left\lfloor \frac{aj}{p} \right\rfloor + p - j' \qquad \text{if } \varepsilon_j = -1.$$

It follows

$$a\sum_j j = p\sum_j \left\lfloor \frac{aj}{p} \right\rfloor + pn + \sum_j \varepsilon_j\, j,$$

the sums being over all $0 < j < \frac{p}{2}$. If a is odd this identity, taken modulo 2, shows that $\sum_j \left\lfloor \frac{aj}{p} \right\rfloor$ and n have the same parity. The theorem follows now from Gauss' criterion. □

Remark. The last theorem can also be formulated slightly differently, which will be useful in a moment. For this note that, for a given $0 < x < \frac{p}{2}$, the quantity $\left\lfloor \frac{ax}{p} \right\rfloor$ equals the number of integers y such that $0 < y < \frac{a}{p}x$. Hence the sum in the preceding theorem equals the number of elements in

$$\Delta := \left\{ (x,y) \in \mathbb{Z}^2 : 0 < x < \frac{p}{2},\ 0 < y < \frac{a}{p}x \right\}.$$

The preceding theorem reads then

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

We can finally prove the famous

Theorem (Quadratic Reciprocity Law)**.** *For any pair p, q of different odd primes, one has*

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

Proof. According to the last remark we have

$$\left(\frac{p}{q}\right) = (-1)^{\#\Delta_1}, \quad \left(\frac{q}{p}\right) = (-1)^{\#\Delta_2},$$

where

$$\Delta_1 = \left\{(x,y) \in \mathbb{Z}^2 : 0 < x < \frac{p}{2},\ 0 < y < \frac{q}{p}x\right\}$$
$$\Delta_2 = \left\{(x,y) \in \mathbb{Z}^2 : 0 < x < \frac{q}{2},\ 0 < y < \frac{p}{q}x\right\}.$$

If we reflect the set Δ_2 at the line $y = x$, i.e. if we apply the map $(x,y) \mapsto (y,x)$, then it becomes

$$\Delta_2' = \left\{(x,y) \in \mathbb{Z}^2 : 0 < y < \frac{q}{2},\ 0 < x < \frac{p}{q}y\right\},$$

which also can be written as

$$\Delta_2' = \left\{(x,y) \in \mathbb{Z}^2 : 0 < x < \frac{p}{2},\ 0 < \frac{q}{p}x < y < \frac{q}{2}\right\}.$$

But $\Delta_1 \cup \Delta_2'$ equals the set of points (x,y) with integer coordinates in $(0, \frac{p}{2}) \times (0, \frac{q}{2})$. The cardinality of this set is $\frac{p-1}{2}\frac{q-1}{2}$, and we notice the quadratic reciprocity law as stated. □

Sometimes it is better to unravel the information encoded by the right hand side of the the reciprocity law by saying

One has $\left(\frac{p}{q}\right) = \left(\frac{q}{p}\right)$ unless $p \equiv q \equiv 3 \bmod 4$, when one has $\left(\frac{q}{p}\right) = -\left(\frac{p}{q}\right)$.

We discuss a few applications of the quadratic reciprocity law. A typical one is the answer to the question for which primes p would a given quadratic equation have solutions modulo p.

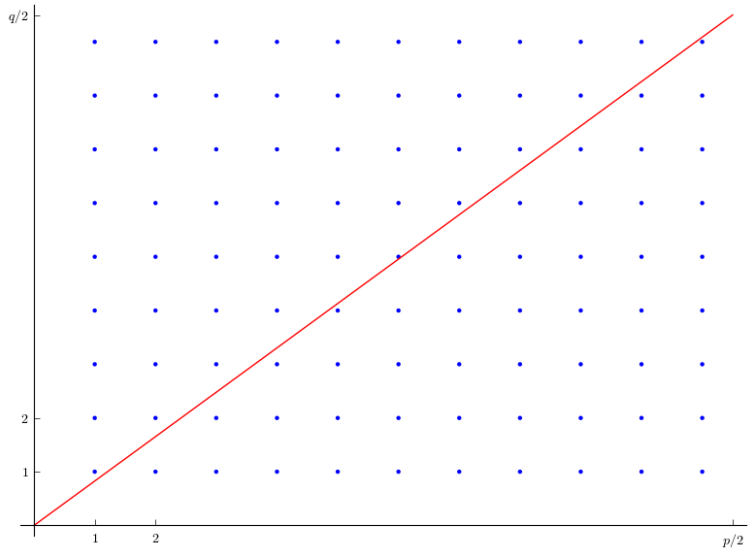


Figure 1. Proof of quadratic reciprocity: By Gauss' criterion we have $\left(\frac{p}{q}\right) = (-1)^{\mathrm{card}(\Delta_1)}$ and $\left(\frac{q}{p}\right) = (-1)^{\mathrm{card}(\Delta_2)}$, where Δ_1 and Δ_2 are the sets of blue points below respectively above the red line $y = \frac{q}{p}x$. Obviously, $\mathrm{card}\,(\Delta_1) + \mathrm{card}\,(\Delta_2) = \frac{p-1}{2} \cdot \frac{q-1}{2}$.

Example. The equation

$$x^2 + x - 1 \equiv 0 \bmod p$$

has discriminant 5. Hence, for $p \neq 2, 5$ it is solvable if $\left(\frac{5}{p}\right) = +1$. But by quadratic reciprocity $\left(\frac{5}{p}\right) = \left(\frac{p}{5}\right)$. The primitive quadratic residues modulo 5 are 1, $2^2 \equiv -1 \bmod 5$, $3^2 \equiv -1 \bmod 5$ and $4^2 \equiv 1 \bmod 5$ (in fact, we could have stopped the calculation after he first two steps since we know that there are not more than two primitive quadratic residues). Therefore, for $p \neq 2, 5$, the given quadratic congruence is solvable if and only if $p \equiv \pm 1 \bmod 5$. A quick check shows that it is not solvable for $p = 2$, and has the solution $x = 2$ for $p = 5$.

Example. Sometimes the quadratic residue symbol allows to produce nice formulas which are useful for further calculations. Such an

example is

$$\#\{0 \le x < p : ax^2 + bx + c \equiv 0 \bmod p\} = 1 + \left(\frac{D}{p}\right),$$

where $D = b^2 - 4ac$. Here a, b, c are given integers and p an odd prime not dividing a. Indeed, the given congruence has no, one or two solutions accordingly as $\left(\frac{D}{p}\right) = -1$, $D \equiv 0 \bmod p$ or $\left(\frac{D}{p}\right) = +1$, respectively.

For applying this formula we ask what the average number ν of solutions of a quadratic congruence modulo p might be. There are $N := (p-1)p^2$ such congruences (as we have $p-1$ choices for a and p choices for b and for c, respectively). The answer to our question is then

$$\nu = \frac{1}{N} \sum_{\substack{a,b,c \bmod p \\ a \not\equiv 0 \bmod p}} \left(1 + \left(\frac{b^2 - 4ac}{p}\right)\right).$$

But for fixed b and a the expression runs with c through all residues modulo p (since a is not divisible by p). Hence

$$\nu = \frac{1}{p} \sum_{D \ bmodp} \left(1 + \left(\frac{D}{p}\right)\right),$$

where we used that $(p-1)p/N = p$. But as there are as many primitive quadratic residues as quadratic non-residues we have $\sum_D \left(\frac{D}{p}\right) = 0$. So we find $\nu = 1$, i.e. the average number of zeros of a given quadratic congruence is 1.

Example. As a last example we answer the question when the cubic congruence $x^2 \equiv a \bmod p$ is solvable for every p. For this note that the application $x + p\mathbb{Z} \mapsto x^3 + p\mathbb{Z}$ defines a map c from $(\mathbb{Z}/p/Z)^*$ to $(\mathbb{Z}/p\mathbb{Z})^*$. Our questions is when is c surjective (since $x^3 \equiv 0 \bmod p$ has anyway always he solution $x \equiv 0 \bmod p$). This is the case if and only if c is injective. But c is a group homomorphism and hence c is injective if and only if the kernel of c is trivial, i.e.!if and only if $x^3 \equiv 1 \bmod p$ has only the solution 1 modulo p. But $x^3 - 1 = (x^2 + x+1)(x-1)$ and hence c is injective if and only if $x^2+x+1 \equiv 0 \bmod p$ has no solution, which is equivalent to $\left(\frac{-3}{p}\right) = -1$ (as $x^2 + x + 1$ has discriminant -3). Finally, by quadratic reciprocity $\left(\frac{-3}{p}\right) = \left(\frac{p}{3}\right)$ (the

reader should verify this). An so the correct answer is: $x^3 \equiv a \bmod p$ is solvable for every a if and only if $p \equiv -1 \bmod 3$ or $p = 3$. Note that we also proved that c is 3 to 1 if $p \equiv 1 \bmod 3$. All this can also be proven by rewriting the equation $x^3 \equiv a \bmod p$ in terms of primitive roots. The condition $p \equiv -1 \bmod 3$ means nothing else but that 3 does not divide the order $p-1$ of a given primitive root modulo p. We leave the details to the reader.

We finally turn to the question how to compute $\left(\frac{a}{p}\right)$ for a given integer a and a given odd prime p. Note that an effective algorithm helps us to answers the question when $x^{\equiv} a \bmod p$ has a solution. However, finding such a solution is a different question and has a different answer, which we shall no pursue here. One possibility for calculating $\left(\frac{a}{p}\right)$ is Euler's criterion, since computing powers can be done roughly with $\log_2 p$ multiplications as we saw in Section 3.5. On the other hand the quadratic reciprocity law suggests that we could proceed as in the Euclidean algorithm. Let r be the rest of a modulo p, apply quadratic reciprocity, let r' be the rest of p mod r, apply quadratic reciprocity, reduce r modulo r' etc. However, the problem is that $\left(\frac{a}{p}\right)$ is only defined for primes p. The idea to decompose a into prime factors q, and apply quadratic reciprocity to each $\left(\frac{q}{p}\right)$ would destroy the efficiency of our planned algorithm since factoring becomes impossible if a is big. Fortunately, our idea still works since factoring, applying quadratic reciprocity leads to a result which can be formulated without the factorization. In other words we can generalize quadratic reciprocity, so to apply it (almost) without any restrictions.

Definition (Generalized Legendre Symbol)**.** We define, for an odd positive integer b and any integer a,

$$\left(\frac{a}{b}\right) = \prod_{p^\beta || b} \left(\frac{a}{p}\right)^\beta .$$

Theorem (Generalized Quadratic Reciprocity Law)**.** *Let a and b be odd positive relatively prime integers. Then one has*

$$\left(\frac{a}{b}\right)\left(\frac{b}{a}\right) = (-1)^{\frac{a-1}{2}\cdot\frac{b-1}{2}}.$$

For the proof we need the following

Lemma. *The application*

$$(\mathbb{Z}/4\mathbb{Z})^* \times (\mathbb{Z}/4\mathbb{Z})^* \to \{\pm 1\}, \quad (a+4\mathbb{Z}, b+4\mathbb{Z}) \mapsto \langle a|b\rangle := (-1)^{\frac{a-1}{2}\cdot\frac{b-1}{2}}$$

is bilinear (i.e. it satisfies $\langle aa'|b\rangle = \langle a|b\rangle\langle a'|b\rangle$ *and* $\langle a|bb'\rangle = \langle a|b\rangle\langle a|b'\rangle$ *for all* a, a', b *and* b'*).*

Remark. Note that the quadratic reciprocity law can be restated in the form

$$\left(\frac{p}{q}\right) = \left(\frac{q}{p}\right)\rangle\langle [p]|[q]\rangle,$$

where $[x] = x + 4\mathbb{Z}$.

Proof of the lemma. We prove $\langle aa'|b\rangle = \langle a|b\rangle\langle a'|b\rangle$. The linearity in the second argument follows then since $\langle a|b\rangle = \langle b|a\rangle$. The symbol $\langle aa'|b\rangle$ equals -1 if and only if $aa' \equiv -1 \bmod 4$ and $b \equiv -1 \bmod 4$, and $aa' \equiv -1 \bmod 4$ if and only if exactly one of a or a' is congruent -1 modulo 4. Therefore $\langle aa'|b\rangle = -1$ of and only if exactly one of the factors $\langle a|b\rangle$ and $\langle a'|b\rangle$ equals -1, which is the claim. □

Proof of the theorem. Decompose $a = p_1 \cdots p_r$ and $b = q_1 \cdots q_s$ into (not necessarily different) primes p_i and q_j. Note that by assumption each p_i is different from all q_j. We then have

$$\begin{aligned}\left(\frac{a}{b}\right) &= \prod_i \left(\frac{p_i}{b}\right) = \prod_i \prod_j \left(\frac{p_i}{q_j}\right) \\ &= \prod_i \prod_j \left(\frac{q_i}{p_j}\right) \langle p_i|q_j\rangle = N \prod_i \prod_j \left(\frac{q_i}{p_j}\right) = N\left(\frac{b}{a}\right)\end{aligned}$$

where $N = \prod_i \prod_j \langle [p_i]|[q_j]\rangle$. Here, for the third identity we applied the quadratic reciprocity law. Finally, by the lemma we have $N = \langle [a]|[b]\rangle$, which proves the theorem. □

For computing the generalized Legendre symbol using quadratic reciprocity and the Euclidean algorithm it is useful to introduce the 2-adic Hilbert symbol. For non-zero integers a and b we set

$$(a,b)_2 := (-1)^{\frac{a'-1}{2}\frac{b'-1}{2}} (-1)^{\alpha\frac{b'^2-1}{8}} (-1)^{\beta\frac{a'^2-1}{8}},$$

where $a = 2^\alpha a'$, and $b = 2^\beta b'$ with integers $\alpha, \beta \geq 0$ and odd integers a' and b'. We leave it to the reader to verify that $(a, b)_2$ is bilinear (i.e. that $(a_1 a_2, b)_2 = (a_1, b)_2 (a_2, b)_2$). Using the theorem about the value of $\left(\frac{2}{p}\right)$ the generalized quadratic reciprocity law now takes the form

$$\left(\frac{a}{b}\right) = \left(\frac{b}{a'}\right) (a, b)_2$$

for any pair of positive integers a and b, where b is odd and a' is the odd part of a.

Using this form of quadratic reciprocity the algorithm for computing $\left(\frac{a}{b}\right)$ for positive odd b would now be as follows. First we compute the 2-adic Hilbert symbol.

Algorithm: Computation of the 2-adic Hilbert symbol

```
def hilbert(a, b):
    """
    Return the 2-adic Hilbert symbol (a,b)
    for integers a and b.
    """
    assert a !=0 and b!=0, 'Error: (%d,%d):
        zero input.'%(a,b)
    s = 0
    while a%2 == 0: a /= 2; s += 1
    t = 0
    while b%2 == 0: b /= 2; t += 1
    ss = -1 if s%2 == 1 and (b%8 == 3 or b
        %8 == 5) else 1
    tt = -1 if t%2 == 1 and (a%8 == 3 or a
        %8 == 5) else 1
    uu = -1 if a%4 == 3 and b%4 == 3 else
        +1
    return ss*tt*uu
```

It is now easy to compute the Legendre symbol using recursion.

Algorithm: Computation of the generalized Legendre symbol

```
def legendre( a, b):
    """
    Return the generalized Legendre symbol
    for integers a and b, where b is odd
        and
    positive.
    """
    if 1 == b: return 1
    a = a%b
    if 0 == a: return 0
    ap = a
    while ap%2 == 0:
        ap /= 2
    return hilbert(a,b) * legendre(b,ap)
```

6. Arithmetical functions

There are several functions $f(n)$ depending on a non-negative integer n which occur naturally in number theory. Examples are the number $\varphi(n)$ of primitive residue classes modulo n, the sum $d(n)$ of the divisors of a number n, the number of primes dividing n (say, including multiplicities) and the like. Many of these functions share properties which are useful in various situations and which we shall study in this section. Though an *arithmetical function* is usually a map $f : n \mapsto f(n)$ which is somehow motivated by arithmetical considerations it is useful to simply adopt the following definition.

Definition. An *arithmetic function* is a map $f : \mathbb{Z}_{\geq 1} \to \mathbb{C}$.

Example. Examples of arithmetic functions are:

(1) The *Euler φ-function* which associates to n the number $\varphi(n)$ of primitive residue classes modulo n,

(2) the *divisor sum* which associates to n the number $d(n)$ of divisors of n,

(3) and, more generally, for any given k, the function σ_k whose values $\sigma_k(n)$ equal the sum of the kth powers of the divisors of n,

(4) the Liouville-function λ, where $\lambda(n) = (-1)^{\Omega(n)}$ and $\Omega(n)$ equals the number of prime factors of n, counted with multiplicities.

Several of these examples are obtained by summing a given simple function over the divisors of n. More formally, we call G the *summatory function of g* if

$$G(n) = \sum_{d|n} g(d).$$

In such an expression we always mean that d runs over the positive divisors of n. In Table 1 the function in the row G is always the summatory function of the one below in the row g.

$G(n)$	$d(n)$	$\sigma(n)$	$\sigma_k(n)$	n
$g(n)$	1	n	n^k	$\varphi(n)$

Table 1. Summatory functions $G(n) = \sum_{d|n} g(d)$

Another observation is that the functions of all the above examples are *multiplicative*, which means that a given function f is not identically 0 and

$$f(mn) = f(m) \cdot f(n) \quad \text{for all } m, n \text{ such that } \gcd(m, n) = 1.$$

We proved this already for Euler's φ-function and it is not hard to verify this for the other examples. However we shall prove in a moment a theorem which makes it easy to recognize this property. An f which satisfies $f(mn) = f(m) \cdot f(n)$ for all m and n without any restriction is called *strongly multiplicative*. The Liouville λ-function is obviously strongly multiplicative. Note that a multiplicative function always

satisfies $f(1) = 1$: indeed $f(1) = f(1)\,f(1)$, and $f(n) = f(1)\,f(n)$; the latter implies that $f(1) \neq 0$ (since f must not be identically 0), and the former then our claim. Also, for a multiplicative function one has

$$f(n) = \prod_{p^\nu \| n} f(p^\nu).$$

A already used before this writing means that the product is to be taken over all maximal prime powers dividing n. Applying this to the function σ_k gives for example

$$\sigma_k(n) = \prod_{p^\nu \| n} (1 + p^k + \cdots + p^{k\nu}) = \prod_{p^\nu \| n} \frac{p^{k(\nu+1)} - 1}{p^k - 1}.$$

6.1. The ring of arithmetic functions. Many of the properties of arithmetic functions are best understood in terms of a natural structure of a ring which one can define on them. As usual we can add two arithmetic functions f and g by defining their sum $f + g$ by $(f + g)(n) = f(n) + g(n)$.

Definition. The *Dirichlet product of to arithmetic functions f and g* is the arithmetic function $f * g$ which is defined by

$$(f * g)(n) = \sum_{d|n} f(d)\, g(n/d) = \sum_{de=n} f(d)\, g(e).$$

(The second sum is over all pairs (d, e) of positive integers such that $de = n$.)

As a first indicator for the usefulness of these definitions note that, for a given arithmetic function f the summatory function F is nothing else but $f * C$, where C is the functions with single value 1 (i.e. $C(n) = 1$ for all n).

Theorem. *The set $\mathfrak{A}$ of arithmetic functions together with the usual addition and the Dirichlet product satisfies the axioms of a commutative ring with neutral element.*

Proof. We have to show that for our addition and Dirichlet product the following properties are satisfied.

(1) $f + (g + h) = (f + g) + h$

(2) $f + g = g + f$

(3) $f + 0 = f$

(4) $f + (-f) = 0$

(5) $f * (g * h) = (f * g) * h$

(6) $f * g = g * f$

(7) $f * \mathbb{E} = f$

(8) $f * (g + h) = f * g + f * h$

Here 0 denotes the function which is identically 0, and $-f$ is the function such that $(-f)(n) = -f(n)$. Finally, $\mathbb{E}$ denotes the function

$$\mathbb{E}(n) = \begin{cases} 1 & \text{if } n = 1, \\ 0 & \text{otherwise.} \end{cases}.$$

We leave it to the reader to verify these properties. Most of them are quickly checked. For (5) we suggest to verify that both sides, evaluated at an argument n, equal

$$\sum_{abc=n} f(a)g(b)h(c),$$

the sum being over all triples (a, b, c) of positive integers such that $abc = n$. □

Though it is not necessary for the understanding of the following it might be helpful for the interested reader to look at the subsequent results with a bit of abstract algebra. Given a ring R one is interested in the group R^* of units in R. By this one means the subset R^* of elements r in R for which there exists an element s in R such that $rs = sr = 1$ (where 1 denotes the multiplicative unit element in R). If r and r' are units then rr' is a unit, and, in fact, R^* possesses the structure of a group with respect to the multiplication in the ring.

Theorem. *Let $f \in \mathfrak{A}$. Then $f \in \mathfrak{A}^*$ (i.e. there exists a g in $\mathfrak{A}$ such that $f * g = \mathbb{E}$) if and only if $f(1) \neq 0$.*

Proof. If $f * g = \mathbb{E}$ for some g, then in particular $f(1)g(1) = 1$, and therefore $f(1) \neq 0$. Assume vice versa that $f(1) \neq 0$. We define g by

induction: set $g(1) = 1/f(1)$ and

$$g(n) = -\frac{1}{f(1)} \sum_{\substack{d|n \\ d<n}} g(d)\, f(n/d).$$

Clearly then $(g * f)(1) = 1$ and $(g * f)(n) = 0$ for $n \geq 2$. □

Theorem. *The set $\mathfrak{M}$ of multiplicative arithmetic functions is a subgroup of $\mathfrak{A}^*$, i.e.:*

(i) *if f is multiplicative, then $f \in \mathfrak{A}^*$,*

(ii) *if f and g are multiplicative, then $f * g$ is multiplicative too,*

(iii) *if f is multiplicative, then its inverse f^{-1} (with respect to the Dirichlet product) is multiplicative.*

Proof. We saw already that a multiplicative function f satisfies $f(1) = 1$, so that, by the preceding theorem, it is invertible.

Assume that f and g are multiplicative and let $h = f * g$. For proving that h is multiplicative let m and n be two positive and relatively prime integers. We leave it to the reader to verify that the application $(d, e) \mapsto de$ defines a bijection

$$\mathfrak{D}(m) \times \mathfrak{D}(n) \xrightarrow{\cong} \mathfrak{D}(mn),$$

where, for an integer l, we use $\mathfrak{D}(l)$ for the set of divisors of l. Using this bijection we find

$$\begin{aligned} h(mn) &= \sum_{t \in \mathfrak{D}(mn)} f(t)\, g(mn/t) = \sum_{(d,e) \in \mathfrak{D}(m) \times \mathfrak{D}(n)} f(de)\, g(mn/de) \\ &= \sum_{(d,e) \in \mathfrak{D}(m) \times \mathfrak{D}(n)} f(d) f(e)\, g(m/d) g(n/e) \\ &= \sum_{d \in \mathfrak{D}(m)} f(d)\, g(m/d) \sum_{e \in \mathfrak{D}(n)} f(e)\, g(n/e) = h(m) h(n). \end{aligned}$$

The proof that f^{-1} is multiplicative if f is multiplicative is similar (use the formula for f^{-1} from the preceding proof). □

Example. From the theorem it is immediate that the *divisor sum functions* $\sigma_k = C * \mathrm{Id}^k$ are multiplicative, since the constant function $C \equiv 1$, the identity function Id and then also $\mathrm{Id}^k : n \mapsto n^k$ are obviously multiplicative.

If G is the summatory function of a given arithmetic function g it is often useful to be able to express g in terms of G. This is indeed always possible. Namely, that G is the summatory function of g means that $G = C * g$ with C denoting the constant function $n \mapsto 1$. Since $C(1) = 1$ we know that C is invertible. Hence $g = C^{-1} * G$. For turning this into a useful formula we need to study C^{-1}, which we shall do now.

Definition (Möbius' μ-function)**.** The arithmetic function μ which is defined by

$$\mu(n) = \begin{cases} 0 & \text{if } n \text{ is not squarefree,} \\ (-1)^r & \text{if } n \text{ is squarefree and the product of } r \text{ primes.} \end{cases}$$

is called the *Möbius μ-function.*

A number n is called *squarefree* if n is not divisible by the square of a prime, or, equivalently, by a perfect square different from 1.

n	1	2	3	4	5	6	7	8	9	10
$\mu(n)$	1	-1	-1	0	-1	1	-1	0	0	1

Table 2. The first values of the Möbius μ-function

Theorem. *For any positive integer n, one has*

$$\sum_{d|n} \mu(d) = \begin{cases} 1 & n = 1, \\ 0 & n > 1 \end{cases}$$

(i.e.. μ is the inverse of the constant function $C \equiv 1$ with respect to the Dirichlet multiplication).

Proof. Set $G(n) = \sum_{d|n} \mu(d)$, i.e. $G = C * \mu$. Clearly $G(1) = 1$. From the definition of μ it is clear that μ is multiplicative, and so is then G too. Hence, for proving $G(n) = 0$ for $n \geq 2$ it suffices to calculate $G(p^r)$ for any given prime power p^r. But from the definition of μ we have $G(p^r) = \mu(1) + \mu(p) = 0$,which proves the theorem. □

We can finally make the above formula $g = C^{-1} * G$ more explicit.

Theorem (Möbius' inversion)**.** *Let G and g be arithmetic functions. Then one has:*

$$\forall n : G(n) = \sum_{d|n} g(d) \quad \textit{if and only if} \quad \forall n : g(n) = \sum_{d|n} \mu(n/d)\, G(d).$$

Proof. This is an immediate consequence of the preceding theorem which can be restated as $\mu * C = \mathbb{E}$, i.e. $\mu = C^{-1}$. Hence $G = g * C$ if and only if $g = G * \mu$, which is the claim. □

Example. 1. We saw in Section 3.4 that $\sum_{d|n} \varphi(d) = n$ for all n. Via Möbius inversion we obtain the formula

$$\varphi(n) = \sum_{d|n} \mu(n/d)\, d.$$

Another type of inversion which one encounters often in number theory is the following. Let f be an arithmetic function and let, for any positive integer n,

$$F(n) = \sum_{\substack{d|n \\ n/d \text{ squarefree}}} f(d).$$

Here the sum is over all positive divisors d of n such that n/d is squarefree. The above equation can be written shorter as $F = f * \chi_{\mathrm{sf}}$, where $\chi_{\mathrm{sf}}(n) = 1$ if n is squarefree and equals 0 otherwise. Since $\chi_{\mathrm{sf}}(1) = 1$ the function χ_{sf} is a unit, and hence $f = \chi_{\mathrm{sf}}^{-1} * F$. We leave it to the reader to prove

Theorem. *one has*

$$\chi_{\mathrm{sf}}^{-1} = \lambda.$$

(Here λ is Liouville's λ-function.)

In other words, the above formula expressing F in terms of f is equivalent to

$$f(n) = \sum_{d|n} F(d)\, \lambda(n/d).$$

6.2. Growth estimates of arithmetical functions. It is clear that, for all integers $n \geq 1$,

$$\varphi(n) \leq n-1, \quad n+1 \leq \sigma_1(n)$$

since among the n residues moulo n at most the ones different from 0 can be relatively prime to n, and since 1 and n are divisors of any given integer $n \geq 1$. Obviously, one cannot improve thesese estimates as one recognizes if one takes for n a prime. More difficult to prove are the following estimates to below, whose proof we do not give here.

Theorem. *For any $n > 2$, one has the following inequalities*[2]

$$\frac{n}{e^{\gamma} \log\log n + 3/\log\log n} < \varphi(n),$$
$$\sigma(n) < n\left(e^{\gamma} \log\log n + 0.6483/\log\log n\right),$$

where γ is Euler's constant.

We shall not prove this theorem. As the reader might guess from the appearance of the logarithm the proof would involve quite a bit of Analysis. What makes these estimates difficult is the erratic jumping of $\varphi(n)$ and $\sigma(n)$. However, as it turns out the jumping behaviors of these functions are quite similar. Namely, their product $\sigma(n)\varphi(n)$ behaves rather smoothly as follows from the following theorem (see also Figure 3).

Theorem. *For $n > 1$ one has*

$$\frac{6}{\pi^2} < \frac{\sigma(n)\varphi(n)}{n^2} < 1.$$

Proof. We use the formulas

$$\sigma(n) = \prod_{p^r \| n} \frac{p^{r+1}-1}{p-1}. \quad \varphi(n) = n \prod_{p|n} \left(1 - \frac{1}{p}\right),$$

from which we deduce

$$\frac{\sigma(n)\varphi(n)}{n^2} = \prod_{p^r \| n} \left(1 - \frac{1}{p^{r+1}}\right).$$

[2]For the first one see [**?**, Thm. 15 and §9], for the second one [**?**].

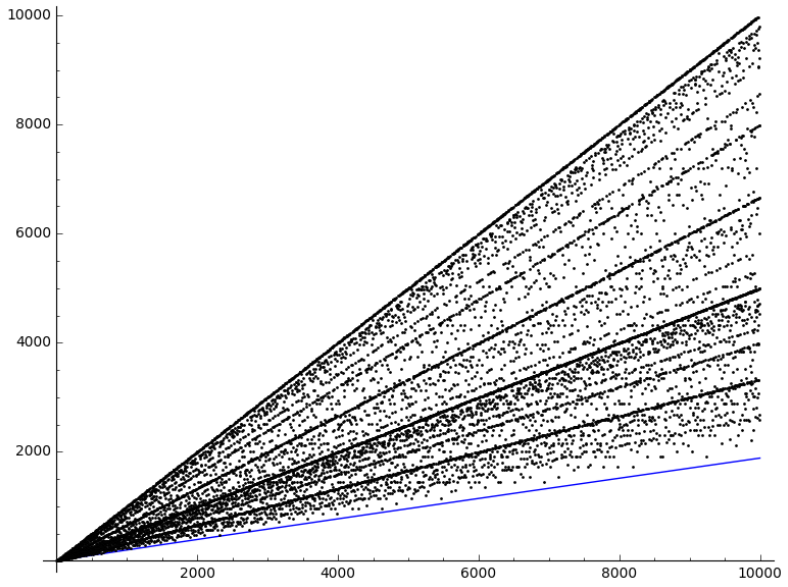


Figure 2. Scatter plot of the points $(n, \varphi(n))$ and the lower bound (in blue) of the theorem. The thick upper bounding line is $y = x - 1$ and is reached by primes n.

The right hand side is (for $n > 1$) obviously strictly smaller than 1 and strictly larger than $\prod_p \left(1 - \frac{1}{p^2}\right)$, where the product is taken over all primes p. We shall prove below that

$$\prod_p \left(1 - \frac{1}{p^2}\right)^{-1} = \sum_{n \geq 1} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

which completes the proof of the theorem. □

The difficulty in getting good estimates for arithmetical functions is that they very often jump in a very irregular manner as we see for example for the number $d(n)$ of divisors of n which falls back to 2 for primes but can also be exponentially big like 2^r if n is the product of r different primes. For obtaining still meaningful estimates it one can try to 'smoothen' the arithmetical function in question. One can try, for example, to estimate the maximum taken over all arguments below a given n, or one can try to estimate the average.

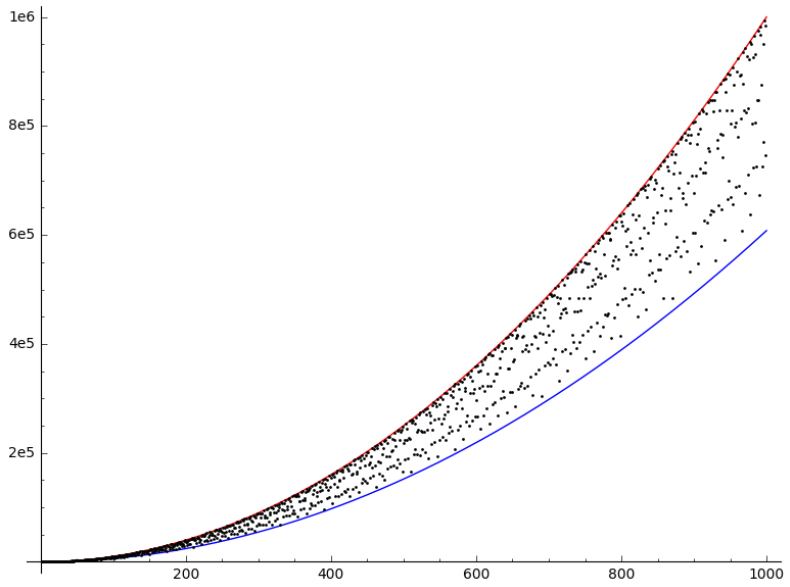


Figure 3. The graphs of $y = x^2$, $y = \frac{6}{\pi^2}x^2$ and the set of points $(n, \sigma(n)\varphi(n))$.

The following two theorems illustrate these ideas. If f and g are two arithmetical function we write $f(n) \sim g(n)$ for indicating that the quotient $f(n)/g(n)$ tends to 0 for increasing n.

Theorem.

$$\max_{k\leq n} d(k) \sim n\frac{\log 2}{\log\log n}$$

We shall not prove this theorem but prove instead the following.

Theorem.

$$\frac{1}{n}\sum_{k=0}^{n}\sigma_1(k) \sim \frac{\pi^2}{12}\,n$$

Proof. This can be seen as follows:

$$\frac{1}{n^2}\sum_{k=0}^{n}\sigma_1(k) = \frac{1}{n^2}\sum_{de\leq n} e = \frac{1}{n^2}\sum_{d\leq n}\sum_{e\leq \frac{n}{d}} e = \frac{1}{n^2}\sum_{d\leq n}\frac{1}{2}\left\lfloor\frac{n}{d}\right\rfloor\left(\left\lfloor\frac{n}{d}\right\rfloor+1\right).$$

Using $\mathrm{fr}(x)$ for the fractional part of a real number x, we can continue

$$= \frac{1}{2}\sum_{d\le n}\frac{1}{d^2} + \frac{1}{2n}\sum_{d\le n}\frac{\mathrm{fr}(n/d)}{d} + \frac{1}{2n^2}\sum_{d\le n}\mathrm{fr}(n/d)\,(\mathrm{fr}(n/d)+1)\,.$$

The second and third sum tend obviously to 0 as n grows. (For this one needs that the harmonic numbers $H_n = \sum_{d\le n}\frac{1}{d}$ satisfy $H_n \sim \log n$.) We finally use a formula that the reader might have seen in some course on calculus (and which we shall prove again later), namely $\sum_{d\ge 1}\frac{1}{d^2} = \frac{\pi^2}{6}$. The theorem becomes now obvious. □

The reader with advanved skills in Analysis might want to try to mimic the proof of the preceding theorem and establish that more generally, for any integer $r \ge 1$.

$$\frac{1}{n}\sum_{k=0}^{n}\sigma_r(k) \sim \frac{\zeta(r+1)}{r+1}\,n^r,$$

where $\zeta(s) = \sum_{n\ge 1}\frac{1}{n^s}$ denotes the Riemann ζ-function (which we shall discuss later), and that

$$\frac{1}{n}\sum_{k=0}^{n}\sigma_0(k) \sim \log n.$$

6.3. Dirichlet series. The formal treatment of arithmetical functions as ring with respect to the Dirichlet product grew naturally out of manipulations of Dirichlet series. In fact, one could present our theory of arithmetical functions as part of the theory of Dirichlet series. However, for reasons of convergence one would have to confine oneself to arithmetic functions of polynomial growth. A third possibility would be to introduce *formal Dirichlet series*. This theory would then perfectly equivalent to our algebraic theory but still notionally very close to the theory of convergent Dirichlet series. The theory of formal Dirichlet series requires a bit of training in algebra though. The interested reader can find it in the appendix to this chapter. In the following paragraphs we explain the basics of the theory of (convergent) Dirichlet series.

Definition (Dirichlet series)**.** A *Dirichlet series* is an infinite series of functions of the form

$$D_f(s) := \sum_{n=1}^{\infty} \frac{f(n)}{n^s}$$

where f is an arithmetical function.

Recall that, for any n and complex number s, one has $n^s = \exp(-s \log n)$. It arises now the natural question if there are real or complex numbers s, for which a given Dirichlet series converges. It is quickly satisfied that a given Dirichlet series converges absolutely for all $\Re(s) \geq \Re(s_0)$ if it converges absolutely for s_0, and it defines then a holomorphic function $D_f(s)$ in the right half plane $\Re(s) > \Re(s_0)$[3] Note also that a Dirichlet series converges absolutely at s_0 if and only if it converges absolutely at every point of the line $\Re(s) = \Re(s_0)$ as follows from the identity $|n^s| = n^{\Re(s)}$ (valid for all real n). The question when a given Dirichlet series converges, say, at at a real k, is easily answered.

Theorem. *The Dirichlet series $D_f(s)$ converges absolutely in some right half plane if and only if f is of polynomial growth (i.e. if and only if for some k the sequence $|f(n)/n^k|$ is bounded for all n).*

Proof. The Dirichlet series

$$\zeta(s) = \sum_{n \geq 1} \frac{1}{n^s}$$

is convergent for $s > 1$. This is quickly checked for example using that, for $n \geq 2$, one has $\frac{1}{n^s} \leq \int_{n-1}^{n} x^{-s}\, dx$, and therefore, for $s > 1$

$$\sum_{n \geq 2} \frac{1}{n^s} \leq \int_1^{\infty} x^{-s}\, dx = s - 1.$$

Assume there for some real C and k one has $|f(n)| \leq Cn^k$ for all n. Then $D_f(s)$ is majorized by $\zeta(s-k)$, hence absolutely convergent for $\Re(s) > k + 1$.

[3]The reader not acquainted to the notion of holomorphic functions may assume in the following that the arguments of the Dirichlet series are real numbers. A Dirichlet series converging absolutely at a real point s_0 converges then absolutely and uniformly in the interval $s \geq s_0$ and defines a continuous function which is analytic function in $s > s_0$.

Assume vice versa that $D_f(s)$ converges at $s = k$ for some real k. Then the sequence $f(n)/n^k$ converges to 0, and, in particular, it is bounded. □

We can now describe the relation between the ring of arithmetic functions and Dirichlet series. For this let $\mathfrak{D}$ denote the set of Dirichlet series which converge absolutely in some right half plane, and let $\mathfrak{A}_{\mathrm{pg}}$ the subset of $\mathfrak{A}$ consisting of all arithmetical functions f which are of polynomial growth, i.e. which satisfy $f(n) = O(n^k)$ for some integer $n \geq 0$.

Lemma. *The set $\mathfrak{A}_{\mathrm{pg}}$ is a subring of $\mathfrak{A}$ (i.e. if f and g are of polynomial growth then $f \pm g$ and $f * g$ are of polynomial growth too).*

Proof. Let f and g be of polynomial growth, say $|f(n) \leq Cn^k$ and $|g(n) \leq C'n^l$ for all n, and let $m = \max(k, l)$. It is clear that $(f \pm g) = O(n^m)$. Moreover,

$$|(f * g)(n)| \leq \sum_{d|n} |f(d)| \cdot |g(n/d)|$$
$$\leq CC' \sum_{d|n} d^k (n/d)^l \leq CC' d(n) n^m.$$

Since $d(n) = O(n)$ we see that $(f * g)(n) = O(n^m + 1)$. □

The following two theorems are straight-forward and left as an exercise to the reader.

Theorem. *The set $\mathfrak{D}$ forms a ring with respect to usual (point-wise) addition and multiplication of complex valued functions. The application $f \mapsto D_f$ defines an isomorphism of rings*

$$\mathfrak{A}_{\mathrm{pg}} \xrightarrow{\cong} \mathfrak{D}.$$

The inverse of a Dirichlet series in $\mathfrak{D}$ is again an element of $\mathfrak{D}$. This statement is equivalent to

Theorem. *One has*

$$\mathfrak{A}^*_{\mathrm{pg}} = \mathfrak{A}^* \cap \mathfrak{A}_{\mathrm{pg}}.$$

The fact that a given arithmetic functions f is multiplicative corresponds to the fact that D_f factors as an Euler product. More precisely, we have

Theorem. *Let f be a multiplicative arithmetic function such that $D_f(s)$ converges absolutely in a right half plane $\Re(s) > k$. Then, for all $\Re(s) > k$, one has*

$$D_f(s) = \prod_p \sum_{l \geq 0} f(p^l)\, p^{-ls},$$

where the product is taken over all primes p, and is absolutely convergent.

Proof. From the assumption it is clear that, for each prime p, the series $\sum_{l\geq 0} f(p^l)\, p^{-ls}$ is absolutely convergent for $\Re(s) > k$. Let $p_1 < p_2 < \dots$ be the series of prime numbers and let $\Re(s) > k$. Then, for any N, one has

$$\prod_{j=1}^{N} \sum_{l\geq 0} f(p_j^l)\, p_j^{-ls} = \sum_{l_1,\dots,l_N \geq 0} \frac{f(p_1^l)\cdots f(p_N^l)}{(p_1^{l_1}\cdots p_N^{l_N})^s} = \sum_{n\in X_N} \frac{f(n)}{n^s},$$

where X_N denotes the set of positive integers containing only the primes p_j $(1 \leq j \leq N)$. It follows

$$\left| D_f(s) - \prod_{j=1}^{N} \sum_{l\geq 0} f(p_j^l)\, p_j^{-ls} \right| \leq \sum_{n>N} \frac{|f(n)|}{n^{\Re(s)}},$$

where we used that every positive integer containing at least one prime p_j with $j \geq N$ is larger than N. By assumption the right hand side of this inequality converges to 0 as N grows. This proves the theorem. □

We finally turn to concrete examples. The simplest non-trivial Dirichlet series is D_C, which we shall according to common notation denote by $\zeta(s)$. Thus, for $\Re(s) > 1$,

$$\zeta(s) = \sum_{n\geq 1} \frac{1}{n^s} = \prod_p \frac{1}{1-p^{-s}}.$$

f	$D_f(s)$
σ_k	$\zeta(s)\zeta(s-k)$
μ	$1/\zeta(s)$
φ	$\zeta(s-1)/\zeta(s)$
λ	$\zeta(2s)/\zeta(s)$

Table 3. Dirichlet series of classical arithmetical functions

Here the second identity follows from the last theorem and $\sum_{l\geq 0} p^{-ls} = 1/(1-p^{-s})$. This identity is (for positive integers s) contributed to Euler and known as *Euler's identity*.

Recall that $f * C$, for any f in $\mathfrak{A}_{\mathrm{pg}}$, is the summatory function of f. Accordingly, we have

$$\zeta(s)\sum_{n\geq 1}\frac{f(n)}{n^s} = \sum_{n\geq 1}\frac{\sum_{d|n} f(d)}{n^s}.$$

Many D_f can be expressed in terms of the Riemann zeta function. The reader can find examples of this in Table 3. For verifying the entries of the table note that $D_{\mathrm{Id}^k}(s) = \zeta(s-k)$ (where as in Section 6.1 $\mathrm{Id}^k : n \mapsto n^k$). Using these identities the first three entries in the table follow from $\sigma_k = C * \mathrm{Id}^k$, $\mu = C^{-1}$, and $C * \varphi = \mathrm{Id}$, respectively. The last entry follows from $C * \lambda = \chi_\square$ and $D_{\chi_\square}(s) = \zeta(2s)$, where $\chi_\square(n) = 1$ if n is a perfect square, and $\chi_\square(n) = 0$ otherwise.

7. Remarks

7.1. Formal Dirichlet series. Identities for arithmetic functions can often be easily recognized and proved using Dirichlet series. For example the identity $\lambda^{-1} = \chi_{\mathrm{sf}}$ from Section 6.1 becomes almost trivial when rewritten in terms of Dirichlet series:

$$D_\lambda(s) = \prod_p \frac{1}{1+p^{-1}},$$

so that $1/D_\lambda(s) = \prod_p 1 + p^{-1}$, which obviously equals $D_{\chi_{\mathrm{sf}}}(s)$, i.e. the $\sum_n 1/n^s$, where n runs over all squarefree integers. Unfortunately, such arguments seem to require a region of absolute convergence for the underlying series and hence apply only to arithmetic functions

of polynomial growth. However, this is not quite true. In fact, it is possible to extend the algebraic theory of arithmetic functions so to mimic as close as possible the desired manipulations with Dirichlet series. This leads to the theory of *formal Dirichlet series*, which we shall now explain. The reader inexperienced in algebra might want to skip this section.

If f is an arithmetic function and z a complex number then zf, i.e. the function $n \mapsto zf(n)$, is also an arithmetic function. This multiplication of arithmetic functions by scalars (together with the addition of arithmetic functions) turns $\mathfrak{A}$ into a vector space over the complex numbers $\mathbb{C}$. In fact, $\mathfrak{A}$ equipped with this scalar multiplication, the addition of functions and the Dirichlet product satisfies the axioms of a *commutative algebra over* $\mathbb{C}$.

Let $\{f_i\}_{i\in I}$ be a (possibly infinite) family[4] of arithmetic functions such that for each integer $n \geq 1$ one has $f_i(n) = 0$ for all but finitely many i in I. We call such a family *summable*. We can then define the *sum of the family* $\{f_i\}_{i\in I}$, denoted by

$$\sum_{i\in I} f_i$$

as the arithmetic function which associates to a given integer $n \geq 1$ the (finite) sum $\sum_{i\in I} f_i(n)$.

Special summable families are obtained as follows. For an integer $n \geq 1$, we use n^{-s} for the arithmetic function which maps n to 1 and $n' \neq n$ to 0. The family $\{n^{-s}\}_{n\in\mathbb{Z}_{\geq 1}}$ is obviously summable, and so is $\{f(n)\, n^{-s}\}_{n\in\mathbb{Z}_{\geq 1}}$ for any given arithmetic function f. Using these notations we can now write every arithmetic function f as sum of the family $\{f(n)\, n^{-s}\}_{n\in\mathbb{Z}_{\geq 1}}$, i.e. we can write any f in the form

$$f = \sum_{n\geq 1} f(n)\, n^{-s}.$$

The expression on the right is called *a formal Dirichlet series*. The reader should note that we did not introduce a new mathematical object, but merely a new language for treating arithmetic functions. We shall see that this language has certain advantages.

[4] A *family* $\{x_i\}_{i\in I}$ *of elements of a set* X is nothings else than a map $I \to X$, $i \mapsto x_i$.

We encourage the reader to practice calculating with formal Dirichlet series and thereby discover natural rules to manipulate them. The first thing to discover is that the Dirichlet product becomes very natural in the language of formal Dirichlet series. Namely, one has $m^{-s} * n^{-s} = (nm)^{-s}$. Because of this one usually uses the dot "$\cdot$" for the operation symbol $*$ when dealing with formal Dirichlet series, and sometimes one simply omits the dot, so that we can write, for example, $m^{-s}n^{-s} = (mn)^{-s}$. The product of two arithmetic functions f and g can then be calculated as

$$\left(\sum_{m\geq 1} f(m)\, m^{-s}\right)\left(\sum_{n\geq 1} g(n)\, n^{-s}\right)$$
$$= \sum_{m,n\geq 1} f(m)g(n)\, m^{-s}n^{-s} = \sum_{l\geq 1}\left(\sum_{mn=l} f(m)g(n)\right) l^{-s}$$

Here we use for the first identity that the product of the sums of two summable families $\{f_i\}_{i\in I}$ and $\{g_j\}_{j\in J}$ equals the sum of the (again summable) family $\{f_i * g_j\}_{(i,j)\in I\times J}$. For the second identity we use that the sum of the family $\{f_i\}_{i\in I}$ equals the sum of $\{\sum_{i\in I_j} f_i\}_{j\in J}$ for any partition of I into a disjoint union of finite sets I_j $(j \in J)$.

We call a (possibly infinite) family $\{f_i\}_{i\in I}$ of arithmetic functions *multiplicable* if we have

(1) for each $n \geq 2$, we have $f_i(n) = 0$ for all but finitely many i,

(2) and $f_i(1) = 1$ for all i.

For a multiplicable family we define the *product* $\prod_{i\in I} f_i$ *of the family* $\{f_i\}_{i\in I}$ as the arithmetic function

$$n \mapsto \sum_{\substack{\{d_i\}_{i\in I}\\ n=\prod_{i\in I} d_i}} \prod_{i\in I} f_i(d_i).$$

Here $\{d_i\}_{i\in I}$ runs through all families of positive integers d_i such that $d_i = 1$ for all but finitely many i. The inner product has to be understood as the finite product over all i in I such that $f_i(d_i) \neq 1$. The sum is over the (finitely) many families $\{d_i\}_{i\in I}$ for which the inner sum is different from zero. The reader should notice that this

product applied to a finite family is nothing else but the Dirichlet product of the members of this family.

Special multiplicable families are obtained as follows: for each prime p let a_p be a map $\mathbb{Z}_{\geq 0} \to \mathbb{C}$ with $a_p(0) = 1$. Then the family

$$\{\sum_{k\geq 0} a_p(k)\, p^{-ks}\}_p$$

(where p runs through all primes) is multiplicable. Indeed, for any given $n \geq 2$, $\left(\sum_{k\geq 0} a_p(k)\, p^{-ks}\right)(n) \neq 0$ for at most one p (namely, at most if n is power of p). The product of such a family is called an *Euler product*. The reader should verify that the product is the arithmetic function

$$n \mapsto \prod_{p^k \| n} a_p(k),$$

where the product is over all prime powers p^k exactly dividing n (i.e. dividing n such that n/p^k is relatively prime to p). The reader should also prove that an arithmetic function f is multiplicative if and only if it can be factored into an Euler product, which means that $f(1) = 1$ and

$$f = \prod_p \sum_{k\geq 0} f(p^k)\, p^{-ks}.$$

It is quickly verified that, for each prime p, the formal Dirichlet series $\sum_{k\geq 0} p^{-ks}$ is the inverse (with respect to Dirichlet multiplication) of $1 - p^{-s}$. Therefore

$$\begin{aligned}
C &= \sum_{n\geq 1} n^{-s} = \prod_p \frac{1}{1-p^{-s}},\\
\mu &= C^{-1} = \prod_p \left(1-p^{-s}\right),\\
\sigma_k &= \sum_{n\geq 1}\sum_{d|n} d^k\, n^{-s} = \prod_p \frac{1}{(1-p^{-s})(1-p^{k-s})},\\
\varphi &= \prod_p \frac{1-p^{-s}}{1-p^{1-s}},\\
\lambda &= \prod_p \frac{1}{1+p^{-s}}.
\end{aligned}$$

Note that from these formulas various identities which we encountered before become rather trivial. For example, the identity $\sum_{d|n} \varphi(d) = n$ reads in terms of formal Dirichlet series $C\varphi = \mathrm{Id}$, which is obvious from the Euler product decomposition of C and φ. Similarly, the identities $\mu = 1/C$ and $\lambda = 1/\chi_{\mathrm{sf}}$ become trivial when rewritten in terms of Euler products.

As an exercise the reader might try to find out how to describe the property of an arithmetic function f to be strongly multiplicative in terms of its Euler product.

7.2. Special values of the Riemann zeta function. We used at several occasions the identity $\zeta(2) = \pi^2/6$. This identity was first proved by Euler in 1734 as a response to the *Basel problem* which asked for the summation of the series $1 + 1/4 + 1/9 + 1/16 + \cdots$. In fact, for every positive even integer the value of $\zeta(s)$ is known to be rational up to a power of π. More precisely, one has the following formula.

Theorem. *For all positive integers k, one has*

$$\zeta(2k) = \pi^{2k} \frac{2^{2k-1}}{(2k)!} |B_{2k}|.$$

Here B_k denotes the kth Bernoulli number.

The Bernoulli numbers are defined by the identity

$$\frac{x}{e^x - 1} = \sum_{k \geq 0} \frac{B_k}{k!} x^k.$$

Multiplying this identity by $e^x - 1$ and comparing coefficients this identity becomes a recursion for the B_k, namely

$$B_0 = 1, \quad \sum_{k=0}^{l-1} \binom{l}{k} B_k = 0 \quad (l \geq 2).$$

Thus $B_1 = -\frac{1}{2}$, $B_2 = \frac{1}{6}$, $B_4 = -\frac{1}{30}$, ... We leave it as an exercise to prove that $B_l = 0$ for all odd $l \geq 3$. This (and many other properties of the Bernoulli numbers) can be easily deduced from their generating series given above.

There is a quick proof that $\zeta(2k) \in \pi^{2k}\mathbb{Q}^*$, which is due to Calabi[5] In the following we explain this proof.

For a positive integer n, let P_n be the polytope in n-dimensional Euclidean space with coordinates $(x_1, \ldots, x_n)$ which is given by the following equations:

$$0 \le x_i \le 1, \quad 0 \le x_i + x_{i+1} \le 1 \qquad (i = 1, \ldots, n).$$

Here we use $x_{n+1} = x_1$. The polytope is bounded (as subset of the unit cube $0 \le x_i \le 1$) and convex (as it is defined as an intersection of half spaces). Let $\mathrm{vol}(P_n)$ the Euclidean volume of P_n. From general theory of poytopes its follows that $\mathrm{vol}(P_n)$ is a rational number since it is defined as intersection of half spaces whose boundaries are hypersurfaces defined by equations with rational coefficients.

Theorem. *For every positive integer k one has*

$$\zeta(2k) = \frac{\pi^{2k}}{2^{2k} - 1} \mathrm{vol}(P_{2k}).$$

Proof. We have

$$\begin{aligned}
\left(1 - 2^{-2k}\right) \zeta(2k) &= \sum_{n=0}^{\infty} \frac{1}{(2n+1)^{2k}} \\
&= \sum_{n=1}^{\infty} \int_0^1 \cdots \int_0^1 u_1^{2n} \cdots u_{2k}^{2n} \, du_1 \wedge \cdots \wedge du_{2k} \\
&= \int_0^1 \cdots \int_0^1 \frac{du_1 \wedge \cdots \wedge du_{2k}}{1 - u_1^2 \cdots u_{2k}^2}.
\end{aligned}$$

We now set

$$u_i = \frac{\sin \frac{\pi}{2} x_i}{\cos \frac{\pi}{2} x_{i+1}} \qquad (1 \le i \le 2k).$$

We leave it to the reader to check that this application maps P_{2k} one-to-one onto the unique cube $[0,1]^{2k}$, and that

$$\frac{du_1 \wedge \cdots \wedge du_{2k}}{1 - u_1^2 \cdots u_{2k}^2} = \left(\frac{\pi}{2}\right)^{2k} dx_1 \wedge \cdots \wedge dx_{2k}.$$

This proves the theorem. □

[5]The second author learned this proof from Calabi 1985 when he was lecturer at the University of Pennsylvania.

Comparing the formulas for $\zeta(2k)$ in terms of Bernoulli numbers and of the preceding theorem we find

$$\mathrm{vol}(P_{2k}) = \frac{2^{2k-1}(2^{2k}-1)}{(2k)!}|B_{2k}|.$$

It would be interesting to have a direct proof of this formula since this would provide, combined with the proof of the previous theorem, a new proof for the well-known formulas for $\zeta(2k)$.

Chapter 3

Primes and factorization

8. Fermat and Mersenne primes

When studying primes and factorization of integers it is natural to investigate numbers of a special shape. In this section we discuss two kinds of such numbers, namely numbers of the form 2^n+1 and 2^n-1. A prime of the form 2^n+1 is called a *Fermat prime*. The first ones are the five numbers $F_k := 2^{2^k}+1$ with $0 \le k \le 4$.

k	0	1	2	3	4
F_k	3	5	17	257	65,537

In fact, not more is currently known. The first observation for their study is the following theorem.

Theorem. *If $p = 2^n + 1$ is a prime number then n is a power of 2.*

Proof. If $n = 2^k m$ with an odd number $m > 1$ then, setting $u = 2^{2^k}$, the number $2^n + 1 = u^m + 1$ factors as

$$u^m + 1 = \big(u + 1\big) \cdot \big(u^{m-1} - u^{m-2} + u^{m-3} - \cdots + 1\big),$$

and both factors are different from 1. □

For an integer $k \ge 0$ the number

$$F_k := 2^{2^k} + 1$$

is called the *kth Fermat number*. A natural question is when F_k is indeed a prime. Since the time of Fermat people tried to factor more and more Fermat numbers and they searched for criteria to recognize the primes among them. The following two theorems give two such criteria.

Theorem (Pépin's test)**.** *Assume $k \geq 2$. Then F_k is a prime number if and only if*

$$3^{\frac{F_k-1}{2}} \equiv -1 \bmod F_k.$$

Proof. Set $N := F_k$. Since $k \geq 2$ we have $N \equiv 2 \bmod 3$ and $N \equiv 1 \bmod 4$. The first congruence implies $\left(\frac{N}{3}\right)$=-1, and the second $\left(\frac{N}{3}\right) = \left(\frac{3}{N}\right)$, hence

$$\left(\frac{3}{N}\right) = -1.$$

If N is prime the claimed congruence of the theorem is therefore nothing else than Euler's criterion (Section 5).

Vice versa, setting $l = 2^k$, the claimed congruence implies

$$3^{2^l} = 3^{N-1} \equiv 1 \bmod N, \quad 3^{2^{l-1}} = 3^{\frac{N-1}{2}} \not\equiv 1 \bmod N,$$

and hence the order of 3 modulo N equals $N-1$. But then N has $N-1$ primitive residue classes (represented by the powers of 3) and therefore N must be prime. □

Theorem. *Assume $k \geq 2$. If p is a prime divisor of F_k, then*

$$p \equiv 1 \bmod 2^{k+2}.$$

Proof. Let p be a prime divisor of F_k. Clearly, $2^{2^k} \equiv -1 \bmod p$, which implies that the order of 2 modulo p is 2^{k+1}, and therefore $2^{k+1} \mid p-1$. Since $k \geq 2$ the latter implies in particular $p \equiv 1 \bmod 8$, and therefore $2 \equiv a^2 \bmod p$ for some a. But this means

$$a^{2^{k+1}} \equiv 2^{2^k} \equiv -1 \bmod p,$$

which implies that the order of a modulo p is 2^{k+2}, which in turn implies the claimed congruence. □

The Fermat primes occur also in another context. Namely, one has the following theorem of Gauss.

Theorem (Gauss)**.** *A regular n-gon can be constructed with compass and straightedge if and only if $\varphi(n)$ is a power of* 2*.*

Constructing a regular n-gon means in a properly chosen coordinate system of the Euclidean plane (which we identify with the complex plane) constructing a primitive nth root of unity ζ[1]. A construction with compass and straightedge consists of exhibiting a sequence $P_0 = 0$, $P_1 = 1$, P_2, ... of points in the plane where each point P_l is the intersection of lines, circles or lines with circles, and where these lines are lines through points P_k and where these circles have as midpoints points P_k and as diameters distances between points P_k, always with $k < l$. In particular, the coordinates of each P_l are solutions of linear equations or quadratic equations whose coefficients are obtained from the coordinates of points P_k $(k < l)$ by applying elementary arithmetic operations. Therefore, constructing ζ means that ζ can be obtained as the last element of a finite sequence of complex numbers where each number is obtained from numbers occurring earlier in the sequence via an elementary arithmetic operation or via taking a square root. The inverse operation to taking a square root is squaring. It is then plausible that the degree of ζ (i.e. the minimal degree which a polynomial with rational coefficients having ζ as zero can have) is a power of 2 if ζ is constructible, and vice versa[2]. On the other hand one knows that this minimal degree is $\varphi(n)$ (namely, the degree of the *nth cyclotomic polynomial* $\prod_{d|n}(x^d - 1)^{\mu(n/d)}$, which is the unique monic polynomial with rational coefficients and minimal degree having ζ as a root). This reasoning can in fact be turned into a rigorous proof of Gauss' theorem using the *theory of field extensions*.

Returning to number theory it is easy to prove the following.

Theorem. *For a positive integer n, the number $\varphi(n)$ is a power of* 2 *if and only if*

$$n = 2^t p_1 p_2 \cdots p_r$$

for some $t \geq 0$ and Fermat primes $p_1 < p_2 < \cdots < p_r$.

[1] This means $\zeta^n = 1$ and n is the smallest positive integer with this property.

[2] In fact, the constructible numbers form a subfield of $\mathbb{C}$ which can be characterized as the field of algebraic numbers whose degree is a power of 2.

Proof. Since φ is multiplicative it suffices to prove the theorem for prime powers $n = p^r$. For $p = 2$ we have $\varphi(2^r) = 2^{r-1}$. If p is odd then $\varphi(p^r) = p^{r-1}(p-1)$ and this is a power of 2 only if $r = 1$ and $p = 2^s + 1$ for some s. □

The only known numbers n for which an n-gon can be constructed are therefore the numbers of the form

$$n = 2^t \prod_{p \in F} p,$$

where F is one of the 2^5 subsets of the five known Fermat primes. The first of these numbers are $= 3, 4, 5, 6, 8, 10, 12, 15, 16, 17, 20, 24$. The 65537-gon was explicitly constructed by Johann Gustav Hermes in 1894 after 10 years of work[3]

Another kind of special primes are the *Mersenne primes*, which are primes of the form

$$M_n := 2^n - 1$$

for some integer $n \geq 1$. The numbers M_n, without the requirement to be prime, are called Mersenne numbers, named after Marin Mersenne who studied the Mersenne primes in the early 17th century. The first Mersenne primes are

n	2	3	5	7	13	17	19	31
M_n	3	7	31	127	8,191	131,071	524,287	2,147,483,647

The largest known one (as of June 2016) is

$$2^{74,207,281} - 1 = 153634\ldots673003 \qquad (22,338,618 \text{ decimal digits})$$

The above table suggests the following observation.

Theorem. *If M_n is a prime than p is prime.*

Proof. If $n = ab$ with $a, b > 1$ then M_n factors as

$$M_{ab} = \big(2^a - 1\big) \cdot \big((2^a)^{b-1} + (2^a)^{b-2} + \cdots + 1\big),$$

which proves the claim. □

[3]The notes for the construction can still be consulted in the mathematics library of the university of Göttingen. Hermes submitted his construction as doctoral thesis to the university and obtained indeed his degree with the support of Felix Klein.

For testing whether a Mersenne number is a prime one usually applies the following theorem.

Theorem (Lucas-Lehmer test)**.** *Let p be an odd prime. Then M_p is a prime of and only if M_p divides the term C_{p-1} of the sequence defined by*

$$C_1 = 4, \quad C_n = C_{n-1}^2 - 2 \quad (n \geq 2)$$

The first six terms of the *Lucas-Lehmer sequence* C_k are

$$4 \quad 14 \quad 194 \quad 37,634 \quad 1,416,317,954 \quad 2,005,956,546,822,746,114$$

Proof of the theorem. Let l be a prime, and let $\mathcal{A}$ be the ring of matrices of the form $\left[\begin{smallmatrix} a & 3b \\ b & a \end{smallmatrix}\right]$, where a and b are in $\mathbb{F}_l$. The word ring indicates that $\mathcal{A}$ is closed under addition, differences and multiplication of matrices (as the reader should verify)[4]. A matrix is $\left[\begin{smallmatrix} a & 3b \\ b & a \end{smallmatrix}\right]$ is invertible if its determinant is different from zero. Let $\mathcal{A}^*$ denote the group of invertible matrices in $\mathcal{A}$. Since $\mathcal{A}$ has l^2 elements and the zero matrix is not invertible, we conclude $\operatorname{card}(\mathcal{A}^*) \leq l^2 - 1$. The order of an invertible matrix M in $\mathcal{A}^*$ is by definition the smallest positive integer n such that $M^n = 1$ (where we use 1 for the unit matrix). For the proof of the theorem we need a result from basic group theory, namely that the order n of M satisfies $n \leq \operatorname{card}(\mathcal{A}^*)$. We set

$$A = \begin{bmatrix} 2 & 3 \\ 1 & 2 \end{bmatrix}, \quad B = A^{-1} = \begin{bmatrix} 2 & -3 \\ -1 & 2 \end{bmatrix}.$$

With $[x]$ denoting the residue class of x modulo l we have $A^{2^{n-1}} + B^{2^{n-1}} = [C_n] \cdot 1$ as is quickly verified by induction over n. Multiplying this identity by $A^{2^{n-1}}$, we have equivalently

$$A^{2^n} = [C_n] \cdot A^{2^{n-1}} - 1.$$

Suppose now that $M_p \mid C_{p-1}$. Choosing for l a prime divisor of M_p we find $[C_{p-1}] = 0$ and hence $A^{2^{p-1}} = -1$, which implies that the order of A equals 2^p. But then $2^p \leq \operatorname{card}(\mathcal{A}^*) \leq l^2 - 1$, and hence $M_p < l^2$. This shows that M_p is prime (since otherwise it would possess at least one prime divisor l with $l^2 \leq M_p$).

[4]The educated reader might notice that $\mathcal{A}$ is nothing else than the quotient $\mathbb{F}_l[X]/(x^2 - 3)$ of the polynomial ring $\mathbb{F}_l[X]$ by the ideal generated by $x^2 - 3$. In other words, $\mathcal{A}$ equals the field with l^2 elements or the direct product of $\mathbb{F}_l$ with itself.

Vice versa, assume M_p is prime. Choose now $l = M_p$. We have to show that $A^{2^{p-1}} = -1$. For this we note $A = \frac{1}{6}C^2$ with $C = \left[\begin{smallmatrix} 3 & 3 \\ 1 & 3 \end{smallmatrix}\right]$. Accordingly

$$A^{2^{p-1}} = A^{\frac{l+1}{2}} = \frac{C^{l+1}}{6^{\frac{l+1}{2}}} = \left(\frac{6}{l}\right)\frac{C^{l+1}}{6},$$

where for the last equality we used Euler's criterion (Section 5). But 3 is a quadratic non-residue mod l since $\left(\frac{3}{l}\right) = -\left(\frac{l}{3}\right)$ and $l = 2^p - 1 \equiv 1 \bmod 3$. Similarly, 2 is a quadratic residue modulo l since $l = 2^p - 1 \equiv -1 \bmod 8$. Hence $\left(\frac{6}{l}\right) = -1$. Moreover,

$$C^l = (3 + \left[\begin{smallmatrix} 0 & 3 \\ 1 & 0 \end{smallmatrix}\right])^l = 3^l + \left[\begin{smallmatrix} 0 & 3 \\ 1 & 0 \end{smallmatrix}\right]^l = 3 + 3^{\frac{l-1}{2}}\left[\begin{smallmatrix} 0 & 3 \\ 1 & 0 \end{smallmatrix}\right] = \left[\begin{smallmatrix} 3 & -3 \\ -1 & 3 \end{smallmatrix}\right],$$

where we used $\left[\begin{smallmatrix} 0 & 3 \\ 1 & 0 \end{smallmatrix}\right]^2 = 3$ and again Fermat's little theorem and Euler's criterion. It follows $C^{l+1} = \left[\begin{smallmatrix} 3 & -3 \\ -1 & 3 \end{smallmatrix}\right] C = 6$, and the claim is now obvious. □

The Lucas-Lehmer test is quickly turned into an algorithm.

Algorithm: Luca-Lehmer test

```
def LLt( p):
    """
    Return True are False accordingly as
    2^p-1 is a prime or not.
    """
    M = 2**p-1
    C = Mod(4, M)
    for n in range(2,p):
        C = C**2-2
    return 0 == C
```

The Mersenne primes are connected with a problem from ancient times. A number n is *perfect* if it equals the sum of its positive divisors different from n. The first perfect number are

$$6 \quad 28 \quad 496 \quad 8128 \quad 33,550,336 \quad 8,589,869,056.$$

Theorem (Euler)**.** *An even positive integer n is perfect if and only if*

$$n = 2^{p-1}(2^p - 1)$$

with a Mersenne prime $2^p - 1$.

Proof. If n is of the given form, then

$$\sigma(n) = \sigma(2^{p-1})\sigma(2^p - 1) = (1 + 2 + \cdots + 2^{p-1}) \cdot 2^p = 2n,$$

where $\sigma(n)$ is the sum of divisors of n. Hence n is perfect.

Vice versa, write $n = 2^{p-1}a$ with an integer $p \geq 2$ (recall that n is even) and an odd integer a. Then $\sigma(n) = 2n$ becomes

$$2^p a = (2^p - 1)\,\sigma(a).$$

It follows $(2^p - 1) \mid a$, say $a = (2^p - 1)b$, and hence

$$2^p(2^p - 1)b = (2^p - 1)\,\sigma((2^p - 1)b) \geq (2^p - 1)(b + (2^p - 1)b).$$

But this inequality is obviously possible only as equality, i.e. only for $b = 1$, which in turn implies $\sigma(2^p - 1) = 2^p$, i.e. that $2^p - 1$ is a prime. This proves the theorem. □

We do not know whether there exist odd perfect numbers, nor do we know if there are infinitely many even ones, i.e. if there are infinitely many Mersenne primes. Even more, it is also unknown whether there are infinitely many composite Mersenne numbers M_p with a prime p. In this context the following theorem is interesting;

Theorem (Euler)**.** *Let p be a prime, $p \equiv 3 \bmod 4$. Then $2p + 1$ is prime if and only if $2p + 1$ divides $2^p - 1$.*

Proof. If $2p + 1$ is prime than by Euler's crierion $2^p \equiv \left(\frac{2}{2p+1}\right) \bmod 2p + 1$. By assumption $p \equiv 3 \bmod 4$ and hence $2p + 1 \equiv 7 \bmod 8$, so that $\left(\frac{2}{2p+1}\right) = -1$.

Vice versa, assume $2^p \equiv -1 \bmod 2p + 1$. Let l be a prime divisor of $2p + 1$. The congruence taken modulo l implies that the order of 2 modulo l equals p, and hence $p \mid l - 1$, i.e. $l = pt + 1$ for some integer $t > 1$. Since $l \leq 2p+1$ we conclude $t = 2$ (whence $l = 2p+1$ is prime) or $t = 1$. But the latter means $l = p+1$, which is impossible since this together with $l \mid 2p+1$ would imply $l = p+1 \mid p$, a contradiction. □

A prime p such that $2p+1$ is a prime too is called *Sophie Germain prime*, the first ones of which below 1000 are

$$2, 3, 5, 11, 23, 29, 41, 53, 83, 89, 113, 131, 173, 179, 191, 233, 239, 251,$$
$$281, 293, 359, 419, 431, 443, 491, 509, 593, 641, 653, 659, 683, 719,$$
$$743, 761, 809, 911, 953,$$

It is not known whether there are infinitely many such primes. But if there were infinitely many Sophie Germain primes $p \equiv 3 \bmod 4$ then we would know that there are infinitely many composite Meresenne numbers M_p with a prime p.

9. Factorization

Given a positive integer n we want to know

(1) whether n a prime,

and if not,

(2) its factorization into primes.

There are various algorithms for answering these questions. However, for large n their running time can become too long so that they might not given an answer in realistic time. One measures the performance of an algorithm by giving an upper bound for the running time as function of $\log_2 n$, i.e. the number of digits of of the binary expansion of n, and sometimes in addition by an upper bound in $\log_2 n$ of the storage space needed to run the algorithm.

It is known that the question (1) can be solved in polynomial time. That means that there is an algorithm whose running time is $O\,(f(\log n))$ for some polynomial f. This is the *Agrawai-Kayal-Sayema algorithm* which was found 2002 by the name three computer scientists. The running time of the original algorithm was $O(\log^{12+\varepsilon} n)$ for every $\varepsilon > 0$[5] Shortly afterwards several improvements to this algorithm were made which reduced the running time further. However, it should be noted that there are various other

[5]We use here the Landau O-notation $f(n) = O\,(g(n))$ for indicating that there is a constant C such that $|f(n)| \leq C\,|g(n)|$ for all n. The ε enters since actually the running time of the Agrawai-Kayal-Sayema algorithm is $O\big(\log^{12}(n)\,\log^{k}(\log(n))\big)$ for some $k > 0$.

primality tests which, for small n, perform better than he Agrawai-Kayal-Sayema algorithm.

Finally, there are also non-deterministic primality tests like for example the *Fermat primality test*. One picks randomly integers a relative prime to the given n and checks whether $a^{n-1} \equiv 1 \bmod n$. If we find an a for which this congruence does not hold we know that n must be composite. However, even if all integers below n would pass this test, we cannot be sure that n is a prime: there are composite Numbers such that $a^{n-1} \equiv 1 \bmod n$ holds for all a relatively prime to n, these re the so-called *Carmichael numbers* after Robert Carmichael who found 1910 the first one (after their basic properties had been observed before); the first ones are 561, 1105, 1729. It is even know that there are infinitely many Carmichael numbers (Alford, Granville and Pomerance, 1994).

In this section we discuss three factorization algorithms, i.e. algorithms which find a proper divisor of a given composite n. Iterating this algorithm would lead then to a complete factorization into prime powers.

9.1. Trial division. The most naive algorithm for factoring a given number n is to try sequentially for divisibility by the consecutive primes. This is reasonable if n possesses small prime factors. However, in the worst case n might be the product of two primes of approximately equal size, and then we would need approximately $\pi(n^{1/2})$ divisions before we encounter a non-trivial divisor of n, where $\pi(x)$ is the number of primes $\leq x$. If $b(n)$ denotes the number of decimal digits of n, so that $\log n^{1/2} \approx \frac{1}{2} b(n) \log(10) \approx 1.15 b(n)$ then, by the prime number theorem

$$\pi(n^{1/2}) \approx \frac{\exp\left(1.15 b(n)\right)}{1.15 b(n)}.$$

For a number n with 30 decimal which is the product of two primes of approximately equal size, we would need 2.8×10^{13} trial divisions to identify these two primes.

9.2. The quadratic sieve. Let a positive integer n be given. Assume that n is composite and odd. If $n = de$ with non-trivial divisors d, e, then

$$n = \left(\frac{d+e}{2}\right)^2 - \left(\frac{d-e}{2}\right)^2,$$

and $\frac{d+e}{2}$ and $\frac{d-e}{2}$ are integers since d and e must be odd. In other words, the application $d \mapsto (\frac{d+e}{2}, \frac{-d+e}{2})$ defines a bijection from the set of divisors d of n with $d \leq n/d$ onto the set of pairs of non-negative integers (a, b) with $n = a^2 - b^2$. Hence we could sequentially try for all integers $a \geq \lceil n^{1/2} \rceil$ whether $a^2 - n = b^2$ for a positive integer b and if $\gcd(a+b, n)$ or $\gcd(a-b, n)$ is a non-trivial divisor of n. (Note that $\gcd(a+b, n)\gcd(a-b, n) = n$ since $n = a^2 - b^2$ so that it suffices to check one of $\gcd(a+b, n)$ or $\gcd(a-b, n)$ for being a proper divisor of n). Eventually we shall find a proper divisor of n. For a divisor d of n, we have $\frac{d+e}{2} \leq \frac{1+n}{2}$ since the function $y = x + \frac{n}{x}$ is decreasing between $x = 1$ and $x = \sqrt{n}$. (In particular, if we do not know whether n is composite we can stop as soon as $a > \frac{1+n}{2}$ to be sure that n is a prime). The worst case for our algorithm would be that n is a product of a very small and a very big prime, so that the first pair $n = a^2 - b^2$ would be for $a \approx \frac{1+n}{2}$ so that we would need $\approx \frac{1+n}{2} - \sqrt{n}$ steps to find the factorization of n. This is much worse than trial division. On the other hand, if n is a product of two primes $p < q$ with, say $\Delta := q - p < \varepsilon\sqrt{n}$ for some small ε, than our algorithm stops successfully at

$$a = \frac{p+q}{2} = \left(n + \left(\frac{p-q}{2}\right)^2\right)^{1/2}$$
$$= \left(n + \left(\frac{\varepsilon\sqrt{n}}{2}\right)^2\right)^{1/2} \approx \sqrt{n}\left(1 + \frac{\varepsilon^2}{8}\right),$$

i.e. already after $\approx \frac{\varepsilon}{8}\Delta$ steps.

However, we can improve the above algorithm. Suppose we have computed $k_j := a_j^2 - n$ for $j = 1, \ldots, r$ and none of these numbers was a square (or at least not a square which led to a proper divisor of n). Then it might still be that a product of some of the k_j is a

perfect square, say $k_{j_1} \cdots k_{j_t} = b^2$. Setting $a := a_{j_1} \cdots a_{j_t}$ it follows

$$a^2 \equiv b^2 \bmod n.$$

It is clear that $\gcd(a+b, n)$ and $\gcd(a-b, n)$ are non-trivial divisors of n unless $a \equiv \pm b \bmod n$.

The natural question is how to ensure that we can find a product of numbers k_j which is a perfect square. A hint is given by the following observation.

Theorem. *Let $B \geq 2$, and let $\pi(B)+1$ positive integers be given all of whose prime divisors are $\leq B$. Then there exists a subset of these numbers whose product is a perfect square.*

Proof. Let $k = \pi(B)$ and $p_1 < p_2 < \cdots < p_k$ be the primes less than or equal to $\leq B$. To a positive integer a which only contains the primes p_j, say $a = p_1^{e_1} \cdots p_k^{e_k}$, we associate its *exponent vector mod 2* $v(a) := (e_1 + 2\mathbb{Z}, \ldots, e_k + 2\mathbb{Z})$ in $\mathbb{F}_2^k$. This application is a homomorphism, i.e. $v(ab) = v(a) + v(b)$. In particular, a is a perfect square if and only if $v(a) = 0$.

The theorem follows now from elementary linear algebra. The $k+1$ exponent vectors mod 2 of the given numbers cannot be linearly independent, i.e. there is a linear combination not all coefficients 0 which is 0. In other words (since we work over $\mathbb{F}_2$) there is subset of these vectors which sums up to the zero vector, and so the product of the corresponding numbers is a perfect square. □

A positive integer having only prime divisors $\leq B$ is called *B-smooth*. Using the last theorem we can set up the following algorithm for factoring:

(1) Choose a B, and examine the numbers $x^2 - n$ where x runs through the integers $\geq \lceil \sqrt{n} \rceil$.

(2) When we have more than $\pi(B)$ numbers $x^2 - n$ which are B-smooth, form their exponent vectors mod 2 and use linear algebra over $\mathbb{F}_2$ to find a non-empty subset S summing up to the zero vector. Form the product of the $x^2 - n$ corresponding to the vectors in S. This product is a perfect square,

say b^2. Let a denote the product of the x in the $x^2 - n$ with exponent vector mod 2 in S. We have $a^2 \equiv b^2 \bmod n$.

(3) If $a \not\equiv \pm b \bmod n$ then output $\gcd(a + b, n)$.

(4) Otherwise return to Step (2), find new subsets S, and if this does not lead to a proper divisor, return to (1) for finding new B-smooth $x^2 - n$ and attempt once again Step (2) and (3).

There are various algorithmic issues: (i) what is the best choice for B, (ii) how can we effectively find subsets S as in (2), and (iii) how can we effectively recognize B-smooth numbers. We shall not discuss in depth the first and second problem. But note that a reasonable choice for B is important. If B is too small then B-smooth $x^2 - n$ will be rare (or do not exist at all) and we have to investigate many more numbers than for a bigger choice. On the other hand a too big B might lead to a long running time for each specimen $x^2 - n$ to be identified as B-smooth or not. For answering the second issue one can, of course, apply Gaussian elimination, but there are also other variants in the literature which might be faster. We discuss the third problem.

For (iii) we can apply successively trial division to the numbers $f(l) = (\lceil \sqrt{n} \rceil + l)^2 - n$ $(l = 0, 1, 2, \dots)$ until we found more than $\pi(B)$ B-smooth numbers. This can be done within a reasonable amount of computing time since we only have to check for prime factors below B, and if we choose B not too big. However, a much faster method is to apply the *quadratic sieve* which gives the name to the whole factoring algorithm. This sieve works as follows.

(1) Create a list $S = [f(0), f(1), \dots, f(N)]$ for some N (whose size will depend on B).

(2) for all primes $p \leq B$ do the following: for all $t = 1, 2, \dots$ solve $f(x) \equiv 0 \bmod p^t$ to obtain all solutions $0 \leq a_j < p^t$ of this congruence. If there is no solution or no solution $a_j \leq N$ restart with the next prime; otherwise replace the entry $S[a_j + p^t k]$ $(k = 0, 1, \dots)$ by $S[a_j + p^t k]/p$.

After the sieve is run the l such that $S[l] = 1$ are exactly those $0 \leq l \leq N$ such that $f(l)$ is B-smooth. There are several improvements

which one could try like for example generating the exponent vectors mod 2 during sieving. A straight forward implementation could look as follows.

Algorithm: Sieving for B-smooth vectors

```
def qs( n, B = 100, N = None):
    """
    Return a list of all B-smooth x^2-n
    among the first N many x>= n^{1/2}.
    """
    w = ceil(sqrt(n))
    f = lambda l: (w+l)^2-n
    if not N: N = prime_pi(B)*100
    S = [ f(l) for l in range(N)]
    for p in primes(B):
        if not Mod(n,p).is_square():
            continue
        t = 1
        while True:
            k = Mod(n, p^t).square_root()
            roots = [(k-w).lift()]
            if p != 2 or (2 == p and 2 == t
                ):
                roots += [ (-k-w).lift()]
            elif t >= 3:
                roots += [ ((1+2^(t-1))*k-w
                    ).lift(), (-(1+2^(t-1))
                    *k-w).lift()]
            if min( roots) >= N: break
            for r in roots:
                k = 0
                while r+k < N:
                    S[r+k] /= p
```

```
                k += p^t
            t += 1
    return [(w+l, f(l)) for l in range(N) if
        S[l] == 1]
```

As stated the quadratic sieve algorithm is deterministic in the sense that, once the parameter B is fixed, it will always output the same result. It might, however, stop without any result. To avoid this one could try to successively increase N. However, even this would not yield necessarily a result. For turning it into a rigorous algorithm (i.e. an algorithm that - apart from running out of time - would in theory eventually yield a factorization) we would have to test all B-smooth products which can be deduced from N, increase B and N if there is no result and restart. By what we saw above we would then (at least) eventually find a solution $n = a^2 - b^2$ with $\gcd(a+b, n)$ being proper divisor of n, but we expect to succeed in general earlier with one of the products formed after the sieving process. In the literature the interested reader can find discussions about the running time of the quadratic sieve.

Example. We apply the quadratic sieve to factor

$$n := F_6 = 2^{64} + 1 = 18,446,744,073,709,551,617.$$

If we run `qs(` $2^{64}+1$`,` $B = 1000$`,` $N = 5120000$`)` we find 240 B-smooth $x^2 - n$ among the first 5,120,000 many

$$x \geq \lceil 2^{64} + 1 \rceil = 9,223,372,036,854,775,809.$$

This is more than enough for a first trial to factor since $\pi(1000) = 168$. The matrix M whose rows are the exponent vectors mod 2 of these $x^2 - n$ has rank 87 and the (left) kernel has therefore dimension 153. This provides us with $2^{153} - 1$ pairs a, b such that $a^2 \equiv b^2 \bmod n$. Picking randomly one of the vectors in the left-kernel of M gives us

$$2,655,324,205,858,794,811^2 \equiv 16,753,529,892,594,067,106^2 \bmod n,$$

and $\gcd(a - b, n) = 274177$. We thus obtain the factorization

$$F_6 = 274,177 \cdot 67,280,421,310,721.$$

Both factors are in fact primes.

9.3. Pollard's ρ algorithm. The second factorization algorithm is particularly suited for composite numbers n which have a small prime factor. It is based on the following idea. Let $f(x)$ be a polynomial with integer coefficients. The natural map $x \mapsto f(x)$ defines a map from $\mathbb{Z}/n\mathbb{Z}$ onto itself. Let us fix for the following an element x of $\mathbb{Z}/n\mathbb{Z}$. Since $Z/n\mathbb{Z}$ is finite there must be integers $0 \leq k < l$ such that $f^k(x) = f^l(x)$. Here f^k means the k-old composition $f \circ \cdots \circ f$ (and f^0 is the identity). It is natural to ask for a non-trivial upper bound for l in terms of n. The following theorem gives us some heuristics for the size of l which we can expect.

Theorem (Birthday problem)**.** *Let x_0, x_1, x_2, ... be uniformly randomly (and independently) chosen from $\{0, 1, \cdots, n-1\}$, and let l be the smallest index such that $x_k = x_l$ for some $k < l$. Then the expected value of l is asymptotically equal to $\sqrt{\pi n/2}$.*

Proof. We have

$$\Pr(l > s) = \big(1 - \frac{1}{n}\big)\big(1 - \frac{2}{n}\big) \cdots \big(1 - \frac{s}{n}\big)$$

for the probability that $l > s$, i.e. that the terms x_0,...,x_s are pairwise different. Moreover,

$$E(l) = \sum_{s=0}^{n} s \Pr(l = s)$$
$$= \sum_{s=0}^{n} s\,[\Pr(l > s-1) - \Pr(l > s)] = \sum_{s=0}^{n-1} \Pr(l > s).$$

Inserting the formula for $\Pr(l > s)$ gives

$$E(l) = \sum_{s=0}^{n-1} \frac{(n-1)(n-2)\cdots(n-s)}{n^s}$$

It can be shown (N-E. Fahssi, 2008, corrected by Vaclav Kotesovec, 2012) that $E(l) \sim \sqrt{\pi n/2}$. □

Suppose p is a non-trivial divisor of n which is small in comparison to n. Denote by $r(x)$ the reduction of x modulo p, and let l' denote the smallest positive integer such that there is a repetition in the sequence

$f^h(r(x))$ $(0 \le h \le l')$. According to the theorem and assuming that the sequences one would therefore expect that

$$l' \approx \sqrt{\pi l'/2} < l \approx \sqrt{\pi l/2}.$$

But then we would have $r\left(f^{l'}(x)\right) = r\left(f^k(x)\right)$ for some $k < l'$, whereas $f^{l'}(x) \neq f^k(x)$. Consequently, $p|d := \gcd(f^{l'}(x) - f^h(x), n) < n$, and so d is a proper divisor of n.

These ideas leads to the following naive algorithm for finding a proper divisor of a given n.

(1) Choose a polynomial f and an initial value x in $\mathbb{Z}/n\mathbb{Z}$.

(2) For each integer $l' = 1, 2, \ldots$ compute $f^{l'}(x)$ and $\gcd(f^{l'}(x) - f^k(x), n)$ for $0 \le k < l'$, and if one these gcds, say d, is different from 1 then stop.

The procedure will eventually stop. If $d = n$, i.e. if the factorization did not succeed then we might try again with different f and x.

Though gcds are quickly computed the number of computed gcds after step l' equals $l'(l'-1)/2$, i.e. it grows quadratically in l'. Moreover, we would have to store all the computed values $f^k(x)$. The following theorem provides an essential improvement.

Theorem (Floyds cycle detection method)**.** *Let $x_0, x_1, \ldots$ be a series which is periodic from a certain index on, and let l be the smallest positive integer such that $x_h = x_l$ for some $0 \le h < l$. Then there exists an $0 < i \le l$ such that $x_i = x_{2i}$.*

Proof. Setting $w = l - h$ we have $x_i = x_j$ for all $i, j \ge h$ such that $i \equiv j \bmod w$. Since w divides $i := l - (h \mod w)$ and $i \ge h$ we have $x_{2i} = x_i$. □

Thus, to recognize that a given sequence becomes eventually periodic it suffices to compare successively the terms x_i and x_{2i}. For the preceding algorithm it means that if there is an l' such that the $\gcd(f^{l'}(x) - f^k(x), n) \neq 1$ for some $0 \le k < l'$, then there is also an $i \le l'$ such that $\gcd(f^{2i}(x) - f^i(x), n) \neq 1$. Namely, $d := \gcd(f^{l'}(x) - f^k(x), n) \neq 1$ means that $f^{l'}(\widetilde{x}) = f^k(\widetilde{x})$, where

$\widetilde{x}$ is the reduction modulo d of x, and then we can apply the preceding theorem which ensures the existence of the claimed i. We hence replace (2) in the above algorithm by

(2)' For each integer $i = 1, 2, \ldots$ compute $d = gcd(f^{2i}(x) - f^i(x), n)$. If $d > 1$ stop.

The algorithm (1) and (2') is the algorithm mentioned in the title of this section (the directed graph consisting of the values $f^i(x)$ as vertices and directed edges from $f^i(x)$ to $f^{i+1}(x)$ looks like a ρ, which explains the word ρ in the algorithm's name).

It is amazingly easy to implement this algorithm, for example as follows.

Algorithm: Pollard's *rho* factorization

```
def prho( n, f = lambda x: x^2+1, iv = 2):
    """
    Return a divisor of n.
    """
    x = y = Mod( iv, n); d = 1; ct = 0
    while 1 == d:
        x = f(x)
        y = f(f(y))
        d = gcd( x - y, n)
        ct += 1
    return ct, d
```

Example. For $n = F_9 = 2^{512} + 1$ (which has 155 decimal digits), `prho( n)` yields the results $(1563, 2424833)$, i.e. it finds the divisor $2,424,833$ of F_9 after $1,563$ iterations.

It can be worthwhile to try other polynomials f. For example `prho( 2^512 + 1, f = lambda x : x^4 + 1)` returns $(425, 2424833)$, i.e. we find a divisor after only 425 iterations, and `prho( 2^512 + 1, f = lambda x : x^8 + 1)` yields even $(174, 2424833)$.

The interested reader can verify that the choice $f(y) = y + 1$ for 

f and $x = 0$ for Pollard's ρ algorithm is nothing else but the trial division algorithm (where we would inspect here, however, successively all numbers and not only primes for being divisors of n).

10. Applications in Cryptography

Many cryptosystems are based on the practical impossibility to factor large integers or to compute general discrete logarithms. Given a group G (like for example the group of primitive residue classes modulo n) the discrete logarithm problem for G asks for an efficient method to solve, for given a and b in the group, the equation $a^x = b$ in x. We shall discuss here the first two of such cryptosystems: RSA and Diffie-Hellman key exchange.

10.1. RSA. The *RSA cryptosystem*, discovered 1977 by Ron Rivest, Adi Shamir, and Leonard Adleman, answers the following question. Is it possible to encipher messages with a publicly known key (and publicly known algorithm) in a way that only the one who issued the public key can decipher it? The practical interest in such a system is immediate. A person issues a public key and he will be the only one who can read those of his received emails which are enciphered with the public key. Or the public key can be used to verify the identity of the issuer by encrypting a message and testing whether the recipient can decipher it.

RSA works as follows. A person A chooses two big primes p and q, computes $n = pq$, and chooses then an exponent e relatively prime to $\varphi(n)$. He publishes then as his *public key*, the pair (n, e). He does not publish the factorization of n or $\varphi(n)$, which are his secrets. A person B who wants to send an encrypted message to A proceeds as follows. He cuts the message (which we might imagine as a sequence of digits to some fixed base) into pieces $m_1, m_2, \dots$ all of which have the same length which is chosen such that $0 \le m_j < n$. He computes than

$$c_j \equiv m_j^e \bmod n, \quad 0 \le c_j < n.$$

The sequence $c_1, c_2, \dots$ is then the encrypted message which he sends to A. A person who wants to decipher this message needs to solve $c_j \equiv x^e \bmod n$ for each j. This is called the *RSA problem*. As we shall see

in a moment the map $()^e : \mathbb{Z}/n\mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ is indeed injective so that $x = m_j$ is the only solution. The security of the RSA cryptosystem is based on the assumption that the most efficient procedure to solve $c_j \equiv x^e \bmod n$ is to calculate $\varphi(n)$, and then d such that $ed \equiv 1 \bmod \varphi(n)$ (recall that e was chosen relatively prime to $\varphi(n)$). Namely then

$$c_j^d \equiv x^{ed} \equiv x \bmod n.$$

The second congruence, for x relatively prime to n, follows on writing $ed = 1 + t\varphi(n)$ for some integer t, and using Euler's theorem which ensures $x^{\varphi(n)} \equiv 1 \bmod n$. We leave it to the reader to show that the congruence $x^{ed} \equiv x \bmod n$ remains true if x is not relatively prime to n. The exponent d is quickly calculated using the extended Euclidean algorithm. However, for this we need $\varphi(n)$, which is kept as a secret by A. Knowing $\varphi(n)$ is essentially equivalent to knowing the factorization of n. Indeed, $\varphi(n) = (p-1)(q-1)$, and if we know $\varphi(n)$ we can calculate p and q as zeros of

$$(y-p)(y-q) = y^2 + (\varphi(n) - n - 1)y + n = 0,$$

i.e. as

$$p, q = \frac{-\varphi(n) + n + 1 \pm \sqrt{(\varphi(n) - n - 1)^2 - 4n}}{2}.$$

10.2. Diffie-Hellman key exchange. This cryptosystem, published 1976 by Whitfield Diffie and Martin Hellman and based on work of Ralph Merkle's to conceptualize public key cryptography, solves the following problem: Is it possible for two persons to agree on a secret key, known only to these two persons, via a public dialogue which can be followed by whoever might be interested? Again the practical impact is immediate: for example encrypted data should be exchanged between two systems which have to agree before the exchange over a public channel on a secret key for en- and deciphering.

The DH key exchange protocol works as follows. Person A and B agree on a prime p and an integer h not divisible by p. Then person A chooses an integers a and B chooses an integer b, respectively, which they do not communicate. But A sends to B the remainder h_A of h^a after division by p, and B sends to A the remainder h_B of h^b after

division by p. The secret key is

$$k \equiv h_A^b \equiv h_B^a \bmod p, \quad 0 \le k < p.$$

A and B are the only persons who can compute h_B^a and h_A^b, respectively since only A knows a and only B knows b. The numbers p, h, h_A and h_B are known by everybody who followed the exchange protocol.

For obtaining the key from the publicly available informations one needs to solve $h^a \equiv h_A$ for a or $h^b \equiv h_B$ for b. In other words, one needs to be able to compute the discrete logarithm x from a given identity $h^x \equiv y \bmod p$. No efficient algorithm is known to compute the discrete logarithm for large p and carefully chosen a. Nevertheless, there are several algorithms which can be used and give quick answers if p is not too big. As example we discuss *Pollard's ρ-algorithm for the discrete logarithm.*

10.3. Pollard's ρ-algorithm for the discrete logarithm. Suppose integer a is relatively prime to N and we want to solve, for a given integer y the congruence

$$y \equiv a^x \bmod N.$$

We assume that such a solution exists (which is for example always the case if a is a primitive root modulo n and y is relatively prime to N). Let n be the order of a modulo N. For solving $y \equiv a^x \bmod N$ we choose a function f from $X := \mathbb{Z}/n\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$ onto itself and an initial value $([u_0], [v_0])$ in X. (We use $[u]$ as abbreviation for the residue class $u + n\mathbb{Z}$.) We compute successively the term of the sequence $([u_k], [v_k]) = f^k([u_0], [v_0])$ of iterates of f until we find an l such that

$$a^{u_k} y^{v_k} \equiv a^{u_l} y^{v_l} \bmod N$$

for some $0 \le k < l$. Replacing y by a^x, it follows

$$u_k + x v_k \equiv u_l + x v_l \bmod n.$$

If this equation is solvable, i.e. if $g := \gcd(v_l - v_k, n)$ divides $\gcd(u_l - u_k, n)$, there will be g solutions, and if g is not too big we can quickly identify the solution x we are looking for. modulo n.

We might hope, for a proper choice of f and the start value $([u_0], [v_0])$, that the sequence $x_k := a^{u_k} y^{v_k}$ is uniformly randomly

(and independently) chosen from the cyclic group of order n generated by a. We can then apply the theorem on the birthday problem from Section 9.3 to conclude that the expected running time for the described algorithm is not worse than approximately $\sqrt{\pi n/2}$. For detecting one of the desired pairs $k < l$ we apply Ford's cycle detection method from Section 9.3. An implementation can now be done as follows.

Algorithm: Pollard's *rho*-algorithm for the discrete logarithm[a]

```
def prhodl( N, a, y, iv = (1,0), F =
    PrhoDL_f, bound = Infinity):
    """
    Return $x$ so tht $a^x \equiv y bmod n$
        .
    """
    a = Mod(a,N); y = Mod(y,N)
    f = F(a,y).map
    n = a.multiplicative_order()
    u,v = iv; u,v = U,V = Mod(u,n), Mod(v,n
        )
    ct = 0

    while ct < bound:
        ct += 1
        u,v = f((u,v))
        U,V = f(f((U,V)))
        if a^u * y^v == a^U * y^V:
            return ct, -(U-u), V-v
    return 'Failed: ct = %d' % ct
```

[a] Note that we implemented the function f here as a method `map()` of a class `PrhoDL_f`. The reason is that the function might need to know about N and a, and we provide this information by creating an appropriate instance of this class after the call of `prhodl`.

The function f is on the original algorithm constructed as follows. One partitions $\mathbb{Z}/N\mathbb{Z}$ into three subsets S_0, S_1 and S_2 of approximately equal size, and defines

$$f([u],[v]) = \begin{cases} ([u],[v+1]) & \text{if } a^u y^v \text{ represents a class in } S_0, \\ ([2u],[2v]) & \text{if } a^u y^v \text{ represents a class in } S_1, \\ ([u+1],[v]) & \text{otherwise.} \end{cases}$$

The partitioning can be achieved by a function $P : \mathbb{Z}/N\mathbb{Z} \to \{0,1,2\}$ and letting S_j be the set of preimages if j. Then $a^u y^v$ represents a class in S_j if $P(r(a^u y^v)) = j$, where r denotes reduction modulo N. For the function P one might take for example the residue class modulo 3 of the smallest positive integer in a given residue class. The above algorithm would then be completed by

Algorithm: A function f for Pollard's *rho*-algorithm for the discrete logarithm

```
class PrhoDL_f:
    """
    Class providing the function for our
    implementation of
    Pollard rho for discete log.
    """
    def __init__( self, a, y):
        self.a = a
        self.y = y
        self.P = lambda c: c.lift()%3

    def map( self, x):
        u,v = x
        j = self.P ( self.a^u * self.y^v)
        if 0 == j: return u,v+1
        if 1 == j: return 2*u,2*v
        if 2 == j: return u+1,v
```

Example. We consider the prime $N = 1,299,709$ and the primitive root $a = 6$. The call `prhodl( N,a,1000)` returns the triple $(1389, 1158192, 216264)$. The equation $1158192 \equiv 216264x \bmod N-1$ has 12 solutions, namely $5907 + t\frac{N-1}{12} \pmod{12}$. For

$$x = 5907 + 6 \cdot \frac{N-1}{12} = 655,761,$$

we find $a^x \equiv 1000 \bmod N$.

It is worthwhile to try different initial values, for example, `prhodl( N,a,1000, iv = (13,1111))` finds x after 440 steps, and `prhodl( N,a,1000, iv = (100,100))` already after 110 steps.

Chapter 4

Diophantine equations

11. Introduction

A diophantine equation is an equation of the form

$$f(x_1, \ldots, x_n) = 0,$$

where $f(a_1, \ldots, x_n)$ is a polynomial in a number n of unknowns with integer coefficients, and where we seek for integral or rational solutions. Many classical problems lead to questions for the solubility or a description of all solutions of diophantine equations.

Example. We want to determine all right triangles whose side lengths are integral or rational. Pythagoras's theorem tells us that our problem is equivalent to solving the diophantine equation

$$a^2 + b^2 = c^2$$

(i.e. the diophantine equation $f(a, b, c) := a^2 + b^2 - c^2 = 0$). The integral solutions of this equation are called *Pythagorean triples*. We shall describe them all in later sections.

Example. We want to determine all natural numbers $n \in \mathbb{Z}_{\geq 1}$ which occur as the area of a right triangle with rational side lengths. Such integers are called *congruent numbers*. By basic theorems from Euclidean geometry a positive integer n is a congruent number if the

following system of diophantine equations is solvable in rational numbers a, b, c:

$$n = \frac{ab}{2}, \quad a^2 + b^2 = c^2.$$

(Note that if this equation has a solution then it has also a solution with a, b, c all three positive).

One can always reduce a system of diophantine equations like the previous one to a single equation. Namely, the system of equations $f_1 = f_2 = \cdots = f_r = 0$ has obviously the same rational or integral solutions as the single equation $f_1^2 + f_2^2 + \cdots + f_r^2 = 0$. However, this is more a theoretical remark. In practice there are better methods to treat systems of diophantine equations. In our case the reader can for example eliminate the variable b by using that $b = 2n/a$ and requiring that the resulting equation $a^4 + 4n^2 = a^2c^2$ has to be solved in rational number a, c. We shall also come back to the congruent number problem in later chapters.

In the following we shall discuss various types of diophantine equations. However before going into details we would like to emphesize that diophantine equations belong to the very heart of mathematics. For the student who encounters them first they might seem to be merely a challenging exercise. However this is far off the truth. We indicate two aspects of this in the following two subsections.

11.1. Hilbert's tenth problem. In a sense every subset of objects in a given set of countably many objects which can be enumerated by some effective procedure (like the subset of prime numbers in the set of all positive integers or the subset of solvable groups in the set of all finite groups) can be encoded by a suitable family of diophantine equations. Thus a huge part of mathematics or computer science is equivalent to the question of solubility of diophantine equations. This is the philosophical interpretation of Matiyasevich's Theorem. This theorem answers also the tenth of Hilbert's 23 problems which Hilbert proposed on the second International Congress of Mathematics, which took place 1900 in Paris, as the outstanding mathematical problems of the coming century.

Hilbert's tenth Problem. *Given a Diophantine equation with any number of unknown quantities and with rational integral numerical coefficients: To devise a process according to which it can be determined in a finite number of operations whether the equation is solvable in rational integers.*

The notion of *process* as Hilbert called it or, as we say today, *algorithm* is meanwhile, after a huge amount of work in fundamental research in mathematics during the first part of the 20th century, rather well understood. This research is connected to names like Gödel, Turing, Neumann, Church and many others[1].

In particular, we have a mathematical precise notion of *recursive sets*. These are those subsets B of the set of non-negative integers $\mathbb{Z}_{\geq 0}$ for which there exists an algorithm which decides for a given integer $n \geq 0$ whether it belongs to B or not. The notion algorithm, and thus recursive set, has been defined in many different ways. For instance we might define B to be recursive by requiring that we can write a program in our favorite programming language which takes as input an integer $n \geq 0$ and outputs True if n belongs to B and False otherwise. However, all definitions have been proven to describe exactly the same family of subsets of $\mathbb{Z}_{\geq 0}$. This led in the end to what is called *Church-Turing thesis*: A function $\mathbb{Z}_{\geq 0} \to \mathbb{Z}_{\geq 0}$ (like the characteristic function of a set B) is computable by a human being ignoring resource limitations if and only if it is computable by a Turing machine.

A related notion is the notion of a *recursive enumerable set*. These are subsets A of $\mathbb{Z}_{\geq 0}$ for which there is an algorithm (e.g. a Turing machine) which enumerates A. We might think of such a set as semi-decidable in the sense that, given an integer $n \geq 0$, we will know after running our algorithm and waiting long enough that n is in A once it is output by the algorithm; on the other hand if we do not get an answer after a certain time we can never be sure that n is not in A. A set B is recursive if and only if B and the complement $\mathbb{Z}_{\geq 0} \setminus B$ are both recursively enumerable.

[1] These mathematical-philosophical considerations which led amongst others to the exact notion of algorithm are in a certain sense the basis of the social changes caused by the increasing digitization of information and automatizing of procedures. On the other hand, fundamental research is also not independent of social changes.

Back to diophantine equations we define

Definition (Diophantine Set)**.** A subset $A \subset \mathbb{Z}_{\geq 0}$ is called *diophantine*, if there exists a polynomial $f(x_1, \ldots, x_r, y)$ with integer coefficients such that

$$A = \{n \in \mathbb{Z}_{\geq 0} \mid \exists x_1, \ldots, x_r \in \mathbb{Z} \colon f(x_1, \ldots, x_r, n) = 0\}\,..$$

Sometimes in the definition of diophantine sets the quantification over all integers is replaced by a quantification over all non-negative integers. However, these two definitions are equivalent. Namely, if a polynomial $f(x_1, \ldots, x_r, y)$ is given we can easily construct a polynomial $g(x_1, \ldots, x_r, y)$ such that, for a given n, the equation $f(x_1, \ldots, x_r, n) = 0$ is solvable in integers y_j if and only if the equation $g(x_1, \ldots, x_r, n) = 0$ is solvable in non-negative integers: for example, one can take $g(x_1, \ldots, x_r, y) = \prod_\varepsilon f(\varepsilon_1 x_1, \ldots, \varepsilon_r x_r, y)$, where ε runs through all vectors of length r with ± 1 as entries. Vice versa, if, for a given $g(x_1, \ldots, x_r, y)$, we set

$$\begin{aligned} f(y_1, z_1, v_1, w_1 \ldots, y_r, z_r, v_r, w_r, y) \\ = g(y_1^2 + z_1^2 + v_1^2 + w_1^2, \ldots, y_r^2 + z_r^2 + v_r^2 + w_r^2, y), \end{aligned}$$

then $f(\ldots, n) = 0$ is solvable in integers if and only if $g(\ldots, n)$ is solvable in non-negative integers. For proving latter equivalence we use Lagrange's four-square theorem, which states that every non-negative integer can be written as sum of four squares.

Another equivalent definition is that a set is diophantine if it equals the set of all non-negative integers assumed by polynomial h with integer coefficients for integer values of its variables. Indeed $n = h(x_i 1, \ldots, x_n)$ if and only if $n - h(x_i, \ldots, x_n) = 0$ is solvable in x_j. On the other hand, if a diophantine set A is given as in the above definition then the condition of solubility of $f(x_1, \ldots, x_r, n) = 0$ in integers x_j is equivalent to

$$\begin{aligned} \exists x_1, \ldots, x_{r+4} \in \mathbb{Z} \colon n = \big(x_{r+1}^2 + x_{r+2}^2 + x_{r+3}^2 + x_{r+4}^2 + 1\big) \times \\ \times \big(1 - f(x_1, \ldots, x_r, x_{r+1} + x_{r+2}^2 + x_{r+3}^2 + x_{r+4}^2 + 1)^2\big) - 1. \end{aligned}$$

It is easy to write a computer program which enumerates a diophantine set. For example we can write a program which tries all possible values for n, x_1,..., x_r, in increasing order of the sum of their

absolute values, and prints n whenever $f(x_1, ..., x_r, n) = 0$. Much more involved is the prove of the inverse statement, whose proof was finally accomplished by Yuri Matiyasevich based on earlier work of Julia Robinson, Martin Davis and Hilary Putnam.

Theorem (Yuri Matiyasevich, 1970)**.** *Every recursive enumerable set is diophantine.*

We encountered already a diophantine set, namely the set of congruent numbers. As we saw this is the set of all positive integers such that $a^4 + 4n^2 - a^2c^2 = 0$ has rational solutions a, c, or, equivalent, such that $a^4+4n^2t^4-a^2c^2 = 0$ has integral solutions a, c, t with $t \neq 0$. (For fulfilling literally the definition of diophantine set the reader may verify that the given conditions on n are equivalent to

$$\exists a, c, x_1, \dots x_4 \in \mathbb{Z}\colon\ a^4 + 4n^2(x_1^2 + x_2^2 + x_3^2 + x_4^2 + 1) - a^2c^2 = 0.$$

(If the x_j run through all integers then $x_1^2 + x_2^2 + x_3^2 + x_4^2 + 1$ runs through all positive integers as follows from Lagrange's four square theorem cited above.) We leave it to the reader to incorporate the additional condition that n should be positive, and to find other examples. Matiyasevich's theorem gives a philosophical reason for this. The reader might want to look up as a non-trivial example a description of the set of prime numbers as diophantine set.

Matiyasevich's theorem provides in fact an answer to Hilbert's tenth problem.

Corollary (Solution of Hilbert's tenth problem)**.** *There is no algorithm which decides, for a given diophantine equation, whether it is solvable or not.*

For deriving the corollary let A be a recursive enumerable but not recursive set. Since A is diophantine we can describe it by a polynomial $f(x_1, \dots, x_r, n)$ as in the definition of diophantine sets. Since A is not recursive there exists no algorithm which decides whether a given n is in A, i.e. whether, for a given n, the diophantine equation $f(x_1, \dots, x_r, n) = 0$ is solvable in integers $x_1, \dots, x_r$. However, it is not a priori clear that there exist recursive enumerable which are not recursive. In mathematical logic it is shown that this is indeed the case.

11.2. Relations to geometry. The second indication for the importance of diophantine equations is that they relate arithmetic and geometry in a deep way which is by far not yet fully explored. If we admit as solutions of the equation $f(x_1, \ldots, x_n) = 0$ real numbers we obtain a geometrical object: a hyper-surface in the affine space $\mathbb{R}^n$. As example consider the equation $x^2 + y^2 = 1$, which defines over the reals the unit circle in the affine plane. It is wise to admit even complex numbers as solutions: we then obtain complex algebraic varieties and we are in the heart of complex algebraic geometry. We shall see that in the case of diophantine equations which define complex curves the topological properties of these curves already determine the qualitative behavior of the set of solutions of the underlying diophantine equation. Finer information is provided by studying in addition the congruences $f(x_1, \cdots, x_n) \equiv 0 \bmod p$ for primes p. We shall get a glimpse of this when discussing Legendre's theorem.

12. Diophantine equations in one variable

Consider a polynomial in one variable with integer coefficients

$$P(x) = a_n x^n + \cdots + a_1 x + a_0.$$

We want to determine all rational solutions of $P(x) = 0$ if there are any. In other words, we look for all integers r and $s \geq 1$ such that $P(r/s) = 0$. We can assume that the fraction r/s is given in its lowest terms, i.e. that $\gcd(r, s) = 1$. We can moreover assume that a_n and a_0 are different from 0 (otherwise omit the terms which are zero and divide by a suitable power of x).

Writing out the equation $P(r/s) = 0$ and multiplying by s^n we obtain

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

But then s divides $a_n r^n$, and since $\gcd(r, s) = 1$, it divides even r. Similarly we note that r divides a_0. We thus have proved

Theorem. *One has*

$$\{r/s \in \mathbb{Q} : \gcd(r, s) = 1,\ s \geq 1,\ P(r/s) = 0\} \subseteq \{r/s \in \mathbb{Q} : r \mid a_0, s \mid a_n,\ s \geq 1\}.$$

We have therefore reduced the problem of solving $P(x) = 0$ in rational numbers to the computation of the values $P(r/s)$ for the finitely many rational numbers r/s described in the theorem. There is an interesting consequence worth to be noted.

Corollary. *Assume that P is monic (i.e. that $a_n = 1$). Then every rational solution of $P(x) = 0$ is integral.*

We note a special case of this:

Corollary. *Let n be a positive integer which is not a perfect square. Then $\sqrt{n}$ is irrational.*

Proof. Indeed, $\sqrt{n}$ is a solution of $x^2 - n = 0$. If it were rational than it would be integral, thus n a perfect square. □

Corollary. *$\sqrt{2}$ is irrational.*

13. Linear diophantine equations

Next we look at diophantine equations in several variable but restrict the degree. We start with the degree 1 case, in other words we want to solve an equation of the form

$$a_1x_1 + \cdots + a_nx_n = b$$

in integers x_j. The coefficients a_j and b are as usual assumed to be integers and the a_j are not all zero. We could also look for rational solutions, but this problem is quickly solved. If, say, a_n is different from zero, then the set of solutions in rational numbers equals the the set of all vectors of the form

$$(x_1, \ldots, x_{n-1}, (b - a_1x_1 + \cdots + a_{n-1}x_{n-1})/a_n),$$

where the x_j are arbitrary rational numbers. If we ask for integral solutions the answer is not so obvious. The next theorem tell us that we can restrict to such equations where the a_j $(j = 1, 2, \ldots, n)$ are relatively prime.

Theorem. *The diophantine equation $a_1x_1 + \cdots + a_nx_n = b$ possesses a solution in integers if and only if $\gcd(a_1, \ldots, a_n)$ divides b.*

The theorem was already proved in Section 1. Namely, the given equation has integer solutions if and only if b is contained in the ideal generated by the numbers a_j, and we have seen that is ideal is generated by the gcd of the a_j.

So we assume that our diophantine equation has solutions. We can then divide the equation by the gcd of the a_j, and the resulting equation has the property that the a_j are relatively prime, which we assume from now on. The question remains how to find and describe the solutions.

Let us assume for a moment that $n = 2$. We have seen in Section 1 that the extended Euclidean algorithm gives generates a solution $x_1^{(0)}$, $x_2^{(0)}$ of $a_1x_1+a_2x_2 = b$. It is now easy to obtain from this all solutions.

Theorem. *The integral solutions of the equation $a_1x_1+a_2x_2 = b$ are given by*

$$x_1 = x_1^{(0)} - ta_2, \quad x_2 = x_2^{(0)} + ta_1,$$

where t runs through the integers.

Proof. If x_1 and x_2 are solutions then $a_1(x_1-x_1^{(0)})+a_2(x_2-x_2^{(0)}) = 0$. It follows that a_2 divides $a_1(x_1 - x_1^{(0)})$, and since a_1 and a_2 are relatively prime, that a_2 divides in fact $x_1-x_1^{(0)}$. Therefore $x_1-x_1^{(0)} = aa_2$ for some integer t. Similarly, $x_2 - x_2^{(0)} = ta_1$ for some integer u. From $a_1(ta_2) + a_2(ua_1) = 0$ we obtain $s = -t$. Vice versa it is clear that any pair x_1, x_2 of the given form provides a solution. □

For extending the last theorem to more tan two variables it is useful to reformulate it in a slightly different form. For this we can assume that $x_1^{(0)} = bu_1$, $x_2^{(0)} = bu_2$ with integers u_1, u_2 satisfying $a_1u_1 + a_2u_2 = 1$. We can write then the general solution of $a_1x_1 + a_2x_2 = b$ in the form

$$(x_1, x_2) = (b, t)\begin{bmatrix} u_1 & u_2 \\ -a_2 & a_1 \end{bmatrix}.$$

The matrix has determinant 1, its inverse is $\left[\begin{smallmatrix} a_1 & -u_2 \\ a_2 & u_1 \end{smallmatrix}\right]$. Vice versa one verifies that for any matrix U with integer entries, determinant 1 and a_1, a_2 in the first column equal, the vector $(b, t)U^{-1}$ runs through all

solutions of our equation when t runs through $\mathbb{Z}$. This is in fact true for any number of variables.

Theorem. *Let U be an $n \times n$-matrix of determinant ± 1 and with $(a_1, \ldots, a_n)'$ as first column*[2]*. Then a vector $(x_1, \ldots, x_n)$ is an integral solution of the equation $a_1x_1 + \cdots + a_nx_n = b$ if and only if it is of the form*

$$(x_1, \ldots, x_n) = (b, t_2, \ldots, t_n)U^{-1},$$

where $t_2, \ldots, t_n$ are integers.

Proof. If $(x_1, \ldots, x_n)$ is of the given form then the x_j are integers. For this we have to verify that U^{-1} has integers as entries. But this follows from the formula $U^{-1} = \det(U)^{-1}U^*$, where U^* is the adjunct of U. Thus, if $(x_1, \ldots, x_n)$ is of the given form it has integral entries and satisfies $(x_1, \ldots, x_n)U = (b, t_2, \ldots, t_n)$, in particular, $(x_1, \ldots, x_n)$ multiplied with the first row of U equals b. But this product is by assumption nothing else but $a_1x_1 + \cdots + a_nx_n$.

Vice versa, if $(x_1, \ldots, x_n)$ is a solution, then $(x_1, \ldots, x_n)U = (b, t_2, \ldots, t_n)$ for suitable integers t_j. □

It remains the question whether such a matrix as in the theorem always exists and how to compute it. For this we note that the $n \times n$-matrices with determinant ± 1 and integral entries form a group, which is usually denoted by $\mathrm{GL}(n, \mathbb{Z})$ with respect to matrix multiplication. This means that for every two matrices U and V in $\mathrm{GL}(n, \mathbb{Z})$ their product and their inverses are in $\mathrm{GL}(n, \mathbb{Z})$ too.

Theorem. *Let a be an integral primitive*[3] *column vector of length n. Then there exists a matrix U in $\mathrm{GL}(n, \mathbb{Z})$ whose first column equals a.*

Proof. The proof will provide also an algorithm for obtaining such a matrix U. In fact we apply the generalized Euclidean algorithm to a for obtaining the desired U.

The extended Euclidean algorithm as explained in Section 2.3 generates a sequence of vectors $a_0 = a, a_1, \ldots, a_k = (1, 0, \ldots, 0)'$

[2] For a vector or matrix X we use X' for the transpose of X.

[3] An integral vector is called *primitive* if its entries are relatively prime.

starting with a and ending with $(1,0,\ldots,0)'$. Each a_{j+1} is obtained from a_j by applying one of the three operations:

(1) Exchange two entries.

(2) Multiply an entry by -1.

(3) Replace the kth entry by the rest after Euclidean division by the lth entry.

But each of these operations corresponds to a matrix multiplication from the left: (1) to multiplication by a permutation matrix, (2) by the diagonal matrix whose diagonal entries are all 1 except for one which is -1, and (3) to a matrix of the form $E + tE_{kl}$, where E is the unit matrix, t an integer and E_{kl} the matrix which has a 1 at the k,lth place and 0 at all others. All these matrices are in $\mathrm{GL}(n,\mathbb{Z})$. Thus we have $a_{j+1} = V_j a_j$ for some V_j in $\mathrm{GL}(n,\mathbb{Z})$, and so $(1,0,\ldots,0)' = V_{k-1}\cdots V_1 V_0 a$. The matrix $U = (V_{k-1}\cdots V_1 V_0)^{-1}$ satisfies then $U(1,0,\ldots,0)' = a$, i.e. it has a as first column. □

It is not hard to transform this algorithm into an algorithm.

Algorithm: Computation of the matrix U[a]

```
def gen_U( a):
    n = len(a)
    A = matrix( ZZ, n, n+1, lambda i,j: a[i
        ] if 0 == j else 1 if j == i+1 else
         0)
    H, U = A.hermite_form( transformation=
        True)
    return U**-1
```

[a] We use here a SAGE method for integer matrices. We form the matrix A which is our primitive vector a followed by the unit matrix to the right. `A.hermite_form( transformation = True)` returns matrices H, V such that H is in Hermite normal form and V is unimodular so that $VA = H$. Since H has first row $(1,0,\ldots,0)'$ the matrix V^{-1} has first row a.

Example. We end this section by an example: we want to determine all solutions of $3x + 5y + 7z = 1$. For this we guess a matrix U

in $\mathrm{GL}(n,\mathbb{Z})$ whose first columns equals $(3,5,7)$. One can take for example

$$U = \begin{bmatrix} 3 & 1 & 0 \\ 5 & 2 & 0 \\ 7 & 0 & 1 \end{bmatrix}$$

The general integral solution of our equation is then

$$\begin{aligned}(x_1,x_2,x_3) = (1,t,u)U^{-1} &= (1,t,u)\begin{bmatrix} 2 & -1 & 0 \\ -5 & 3 & 0 \\ -14 & 7 & 1 \end{bmatrix} \\ &= (2-5t-17u, -1+3t+7u, u),\end{aligned}$$

where t and u are integers.

The reader will have noticed the special shape of U, which is due to the fact that 3 and 5 are relatively prime and can therefore, by Bézout's theorem, be completed to an integral matrix with determinant 1. More generally, if relatively prime integers $a_1,\ldots a_n$ are given, and, say, the first r have already gcd 1 we can find a matrix U' in $\mathrm{GL}(r,\mathbb{Z})$ with first column equals to $(a_1,\ldots,a_r)'$, and then

$$U = \begin{bmatrix} U' & 0_{r,n-r} \\ 0_{n-r.r} & 1_{n-r} \end{bmatrix}$$

(where $O_{r,s}$ denotes the $r \times s$=matrix with all entries 0, and 1_n the $n \times n$-unit matrix) is a matrix in $\mathrm{GL}(n,\mathbb{Z})$ whose first column equals $(a_1,\ldots,a_n)'$.

14. Special quadratic diophantine equations

In this section we discuss diophantine equations of the form $f(x,y) = 0$, where f is a polynomial of degree 2. Recall that this this means that f is a linea combination of monomials $x^k y^l$ with $k+l \leq 2$ and that we have equality for at least one monomial. In fact we shall restrict here to special f, namely $f(x,y) = x^2+y^2-1$, and the family $f_n = x^2 - ny^2 - 1$ (n a positive integer). A complete theory for arbitrary f of degree 2 can be found in Chapter **??** on conic sections. In this section we shall also encounter for the first time the use of geometric arguments to solve diophantine problems.

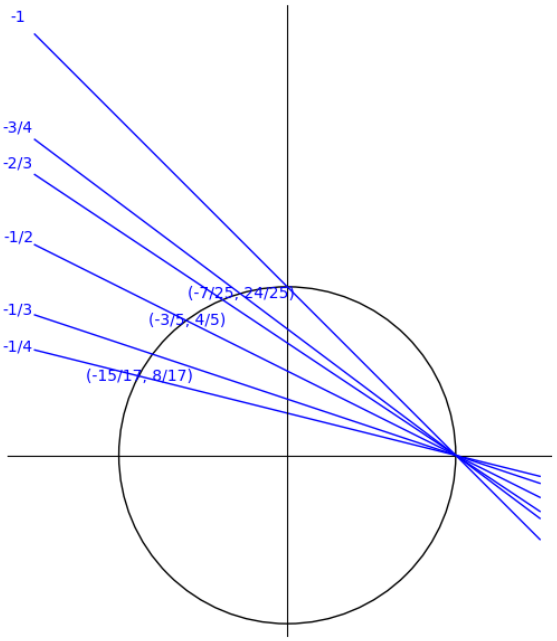


Figure 1. Diophant's method of parametrizing the points on the unit circle by the lines of a pencil.

14.1. Rational points on the unit circle. We consider the diophantine equation

$$x^2 + y^2 = 1.$$

The set of solutions in real numbers is the unit circle in the Euclidean plane. The set of integral solutions consists merely of the four points $(\pm 1, 0)$, $(0, \pm 1)$. However, if we ask for rational solutions the question becomes much more interesting: there are plenty such solutions like e.g. $(3/5, 4/5)$, $(15/17, 8/17)$. In fact, there are infinitely many solutions. The idea to find them all is already describes in a famous book on diophantine problems which is attributed to Diophant who lived around 250 AD.

The idea is to consider the pencil through the point $(1, 0)$ on the unit circle, i.e. the set of lines through this point; see Fig. 1. It is clear that every line in the plane intersects the unit circle in at most

two points. A line intersects the unit circle in only one point if it is a tangent. A line in our pencil through $(1,0)$ which is different from the tangent line (the line perpendicular to the x-axis) intersect the unit circle in exactly one point different from $(1,0)$. Vice versa, every point different from $(1,0)$ lies on exactly one line in our pencil. In other words, the application which associates to every line of the pencil its second intersection point with the unit circle defines a bijection.

The lines of our pencil different from the tangent are given by the equations $y = \lambda(x-1)$, where λ runs trough the reals. As we shall verify in a moment, the slope λ is rational if and only if (x,y) has rational entries. This will therefore solve our diophantine problems.

The intersection point of $y = \lambda(x-1)$ with the unit circle is quickly computed: Elimination of the variable y gives

$$\begin{aligned} &x^2 + \lambda^2(x-1)^2 = 1 \quad (x \neq 1),\\ &x + 1 + \lambda^2(x-1) = 0,\\ &x = \frac{\lambda^2 - 1}{\lambda^2 + 1}, \end{aligned}$$

and inserting back this expression for x in $y = \lambda(x-1)$ then $y = \frac{-2\lambda}{\lambda^2+1}$.

Clearly, if λ is rational so are x and y. Vice versa, if x, y are rational the formula $\lambda = \frac{y}{x-1}$ shows that λ is in $\mathbb{Q}$. We can summarize our reasoning by the following theorem.

Theorem. *One has*

$$\begin{aligned} \big\{(x,y) \in \mathbb{Q}^2 \colon x^2 + y^2 = 1, (x,y) \neq (1,0)\big\}& \\ = \left\{\left(\tfrac{\lambda^2-1}{\lambda^2+1}, \tfrac{-2\lambda}{\lambda^2+1}\right) : \lambda \in \mathbb{Q}\right\}& \end{aligned}$$

Diophant's method can obviously also applied to other quadratic diophantine equations. The reader might try to apply it for example to the diophantine equation $x^2 + xy + y^2 = 0$. However, Diophant's method depends on the knowledge of at least one solution for basing the pencil on it. Such a point does not necessarily exist. In Section 15 we shall give an answer to the question when a solution exist and when not, and how to compute one.

If we look again at our derivation of the equality of the theorem we see that we could apply this method also to other fields, not only

the rational numbers. In particular, we can apply it to the finite fields $\mathbb{Z}/p\mathbb{Z}$, where p is a prime.

14.2. Pythagorean triples. Using the explicit description of the rational points on the unit circle it is now easy to obtain a complete description of all solutions of the equation $a^2 + b^2 = c^2$ in positive integers. Note that it suffices to determine all primitive solutions, i.e. all solutions with relatively prime a, b, c since any solution can be reduced to a primitive one by dividing by its gcd. Also, reducing any primitive solution modulo 4 shows that a or b must be even (since otherwise $1 + 1 \equiv c^2 \bmod 4$, which is not solvable). Obviously we can suppose that, say, b is even.

Theorem (Pythagoraen triples)**.** *The primitive triples a, b, c of positive integers satisfying*

$$a^2 + b^2 = c^2, \quad 2 \mid b$$

are identical with the triples of the form

$$a = p^2 - q^2, \quad b = 2pq, \quad c = p^2 + q^2,$$

where $p > q > 0$ are relative prime integers such that $p + q$ is odd.

Proof. Suppose a, b, c is a primitive Pythagoraen triples with even b. Then $(\frac{a}{c})^2 + (\frac{b}{c})^2 = 1$, and by the theorem of the preceding section we have $\frac{a}{c} = \frac{p^2-q^2}{p^2+q^2}$ and $\frac{b}{c} = \frac{2pq}{p^2+q^2}$ with relatively prime integers p and $q \geq 1$ (we wrote here $\lambda = \frac{p}{q}$). Since $\frac{a}{c}$ is positive we have also $p > q > 0$. Since a, b, c are relatively prime we see that a, c and b, c must be relatively prime, i.e. $\frac{a}{c}$ and $\frac{b}{c}$ are given in lowest terms. From this we deduce that $ta = p^2 - q^2$, $tb = 2pq$, $tc = p^2 + q^2$ for a suitable positive integer t. We show that in fact $t = 1$. Namely, t divides $p^2 - q^2$, $2pq$, $p^2 + q^2$. But then t divides the sum and the difference of the first and last number, i.e. t divides $2p^2$ and $2q^2$. Since p and q are relatively prime we conclude $t = 1$ or $t = 2$. By assumption b is even, and hence a and c are odd. But then, for $t = 2$, from $tc = p^2 + q^2$ we deduce that p and q are odd, hence $2pq$ is exactly divisible by 2, whereas $tb = 2pq$ implies that 4 divides $2pq$, a contradiction. Thus $t = 1$, and since c is odd we deduce from $c = p^2 + q^2$ that $p + q$ is odd.

Vice versa, every triple $a = p^2 - q^2$, $b = 2pq$, $c = p^2 + q^2$ with p and q satisfying the given constraints is obviously a Pythagoraen triples with even b. It is also primitive: if a prime l would divide $p^2 - q^2$, $2pq$, $p^2 + q^2$, it would divide $2p^2$ and $2q^2$, hence would be equal to 2. But $p^2 + q^2$ is odd by assumption, a contradiction. □

a	b	c	a	b	c	a	b	c	a	b	c
3	4	5	11	60	61	33	56	65	55	48	73
5	12	13	13	84	85	35	12	37	63	16	65
7	24	25	15	8	17	39	80	89	65	72	97
9	40	41	21	20	29	45	28	53	77	36	85

Table 1. The 16 primitive Pythagorean triples with even b and $c < 100$

14.3. Pell's equation. The quadratic equation

$$x^2 - ny^2 = 1,$$

where n is a given positive integer is called Pell's equation. Over the reals this equation defines a hyperbola (see Fig. 2). For obtaining the rational solutions x and y we can proceed as in the case of the unit circle in Section 14.1 and apply Diophant's method. If we choose the pencil $y = \lambda(x - 1)$ we obtain

$$\begin{aligned} &x^2 - n\lambda^2(x-1)^2 = 1 \quad (x \neq 1), \\ &x + 1 - n\lambda^2(x-1) = 0, \\ &x = \frac{n\lambda^2 + 1}{n\lambda^2 - 1}, \quad y = \frac{2\lambda}{n\lambda^2 - 1}. \end{aligned}$$

When λ runs through the rational numbers (x, y) runs through the rational solutions different from $(1, 0)$. (If n is a perfect square we have to exclude the two rational solutions of $\lambda^2 n - 1 = 0$.)

However, Pell's equation is a bit more interesting than this. The equation $x^2 + y^2 = 1$ has obviously only four integral solutions: $(1, 0)$, $(0, 1)$, $(-1, 0)$, $(0, -1)$. In contrast to this Pell's equation possesses infinitely many integral solutions as soon as n is not a perfect square. Note that the assumption that n is not a perfect square is necessary

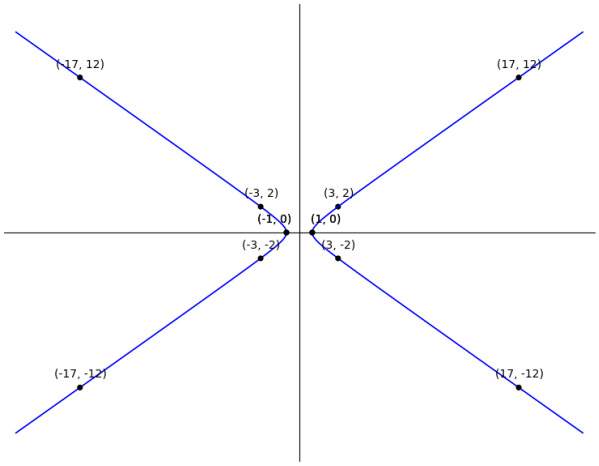


Figure 2. Hyperbola defined by Pell's equation $x^2 - 2y^2 = 1$. The dots represent some integral solutions.

here. We leave it to the reader to show that, for n being a perfect square, say, $n = k^2$, the integral solutions are $(1, 0)$, $(-1, 0)$.

For studying the integral solutions it is useful to represent a solution (x, y) of $x^2 - ny^2 = 1$ by the matrix

$$M := \begin{bmatrix} x & ny \\ y & x \end{bmatrix}.$$

This matrix has determinant 1, which is equivalent to the fact that (x, y) are solutions of Pell's equation. It is quickly checked that the product and the inverses of any two matrices of this form are again of this form, and thus yield new solutions. The 2×2-matrices with integral entries and determinant 1 form a group with respect to matrix multiplication, which is denoted by $\mathrm{SL}(2, \mathbb{Z})$.

The set of matrices in $\mathrm{SL}(2, \mathbb{Z})$ of the above form are in 1 to 1 correspondence to the integral solutions of Pell's equation. They form a subgroup of $\mathrm{SL}(2, \mathbb{Z})$, which we denote by S_n. The group $\mathrm{SL}(2, \mathbb{Z})$ is not commutative, but S_n is, i.e. for any two matrices M and M' in S_n we have $MM' = M'M$. The group S_n contains the matrices $\left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} -1 & 0 \\ 0 & -1 \end{smallmatrix}\right]$, they have the trace $\operatorname{tr} M = \pm 2$. All other matrices M in S_n have non-zero entries and $|\operatorname{tr} M| > 2$. If we can assure that

there is at least one matrix M in S_n different from $\pm\left[\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right]$, then we can conclude that S_n is infinite since all the powers M^k are pairwise different. Namely, $M^k = M^l$ for some $k \geq l$ implies $M^{k-l} = \left[\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right]$. But this is possible only for $k = l$ since the eigenvalues of M are the solutions of $\lambda^2 - t\lambda + 1 = 0$ ($t = \operatorname{tr} M$), i.e. the numbers

$$\lambda = \frac{t \pm \sqrt{t^2 - 4}}{2} = x \pm \sqrt{n}y$$

which cannot be roots of unity if $\operatorname{tr} M^2 > 4$.

Example. We consider the example $x^2 - 2y^2 = 1$. A bit of trying gives us the solution $(3, 2)$. As we saw all powers of $\left[\begin{smallmatrix}3&4\\2&3\end{smallmatrix}\right]$ yield also solutions.

k	1	2	3	4	5
x, y in $\left[\begin{smallmatrix}3&4\\2&3\end{smallmatrix}\right]^k$	$(3,2)$	$(17,12)$	$(99,70)$	$(577,408)$	$(3363,2378)$

In fact, every solution $x, y > 0$ is obtained in this way as follows from succeeding theorem and the obvious fact that there is no solution $x, y > 0$ with $x < 3$.

We shall see in a moment that there exist always positive integral solutions x, y of Pell's equation $x^2 - ny^2 = 1$ (for n not a perfect square), where *positive* means that x and y are both positive. However, we postpone the proof for a moment and prove first of all the following.

Theorem. *Let n be positive and not a perfect square. Let x_0, y_0 be the integral solution of $x^2 - ny^2 = 1$ with smallest x_0 among all positive integral solutions, and let F be its matrix representation. Then the powers F^k ($k \geq 1$) of the matrix F run through all positive integral solutions.*

Proof. For a matrix $M = \left[\begin{smallmatrix}x&ny\\y&x\end{smallmatrix}\right]$ in S_n let $\lambda(M)$ denote the eigenvalue $x + \sqrt{n}y$ of M. It is easily checked that $\lambda(M)\lambda(M') = \lambda(MM')$, that $\lambda(M) \geq 1$ if and only if M has non-negative entries (as follows from $|x| > 1 + \sqrt{n}|y|$???), and that $\lambda(M) = 1$ only for $M = \left[\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right]$. For the matrix F we have $\lambda(F) > 1$. Moreover, $\lambda(F)$ is the minimum of all $\lambda(M) > 1$. Indeed, if, for $M = \left[\begin{smallmatrix}x&ny\\y&x\end{smallmatrix}\right]$ with $\lambda(N) > 1$, we have $x > x_0$ then $ny^2 = x^2 - 1 > x_0^2 - 1 = ny_0^2$, therefore $y \geq y_0$, and then $\lambda(M) > \lambda(F)$.

Therefore, given a matrix M in S_n with positive entries, we can find an integer $k \geq 0$ such that $\lambda(F)^k \leq \lambda(M) < \lambda(F)^{k+1}$, i.e.

$$1 \leq \lambda(F^{-k}N) < \lambda(F).$$

Since $\lambda(F)$ is minimal among all $\lambda(M) > 1$ we conclude $\lambda(F^{-k}M) = 1$, i.e. $M = F^k$. □

If x, y is an arbitrary solution with x, y both nonzero, then exactly one of

$$M := \left[\begin{smallmatrix} x & ny \\ y & x \end{smallmatrix}\right], \quad \left[\begin{smallmatrix} -1 & 0 \\ 0 & -1 \end{smallmatrix}\right] M = \left[\begin{smallmatrix} -x & -ny \\ -y & -x \end{smallmatrix}\right],$$
$$M^{-1} = \left[\begin{smallmatrix} x & -ny \\ -y & x \end{smallmatrix}\right], \quad \left[\begin{smallmatrix} -1 & 0 \\ 0 & -1 \end{smallmatrix}\right] M^{-1} = \left[\begin{smallmatrix} -x & ny \\ y & -x \end{smallmatrix}\right]$$

corresponds to a positive solution. We have therefore

Corollary. *In the notations of the preceding theorem, the matrix of any solution of* $x^2 - ny^2 = 1$ *is of the form* $\left[\begin{smallmatrix} -1 & 0 \\ 0 & -1 \end{smallmatrix}\right]^\nu F^k$ *for suitable integers* $0 \leq \nu \leq 1$ *and* k.

Using the language of group theory one could shorten the previous statement by saying that S_n is the direct group of the cyclic group of order 2 generated by $\left[\begin{smallmatrix} -1 & 0 \\ 0 & -1 \end{smallmatrix}\right]$ and the infinite cyclic group generated by F. We shall henceforth call the positive integral solution with smallest x the *fundamental solution*.

It remains to show the existence of solutions $x, y \neq 0$ and the question how to find the fundamental solution. For this we observe that we can associate to every matrix $A = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ in $\mathrm{GL}(2, \mathbb{R})$[4] the fractional transformation of $\mathbb{P}^1(\mathbb{R}) = \mathbb{R} \cup \{\infty\}$ defined by

$$\xi \mapsto A\xi := \frac{a\xi + b}{c\xi + d}$$

(with the usual conventions $A\infty = \frac{a}{c}$ and $A\xi = \infty$ if $c\xi + d = 0$[5]). We leave it to the reader to verify the following rules. Namely, $\left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ fixes $\mathbb{P}^1(\mathbb{R})$ element-wise (i.e. that $\left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]\xi = \xi$ for every ξ), and that $A(B\xi) = (AB)\xi$. We can use this to characterize the group S_n of solutions of our equation $x^2 - ny^2 = 1$.

[4] $\mathrm{GL}(2, \mathbb{R})$ is the group of all 2×2-matrices with real entries and non-zero determinant.

[5] A less ad-hoc and more conceptual approach to this *action of* $\mathrm{SL}(2, \mathbb{Z})$ *on the projective line* $\mathbb{P}^1(\mathbb{R})$ will be given in Chapter **??**.

Lemma. *Assume that n is not a perfect square. A matrix A in* $\mathrm{SL}(2,\mathbb{Z})$ *is in S_n (i.e. represents a solution of $x^2 - ny^2 = 1$) if and only if $A\sqrt{n} = \sqrt{n}$.*

Proof. Suppose we have, for a given $A = \left[\begin{smallmatrix} x & b \\ y & d \end{smallmatrix}\right]$, the identity $A\sqrt{n} = \sqrt{n}$. Writing out this identity gives, after clearing the denominator,

$$x\sqrt{n} + b = \sqrt{n}\,(y\sqrt{n} + d),$$

from which we infer (using that $\sqrt{n}$ is irrational) $A = [\begin{smallmatrix} x & ny \\ y & x \end{smallmatrix}]$, in particular, that A is in S_n. The inverse implication is obvious. □

Our problem becomes now to find matrices A which leave $\sqrt{n}$ fixed. For this we use the map

$$\phi : \mathbb{P}^1(\mathbb{R}) \to \mathbb{P}^1(\mathbb{R}), \quad \phi(\xi) := \frac{1}{\mathrm{frac}(\xi)},$$

where $\mathrm{frac}(\xi)$ is the fractional part $\xi - \lfloor \xi \rfloor$ of ξ. Note that

$$\xi = a + \frac{1}{\phi(\xi)} = \begin{bmatrix} a & 1 \\ 1 & 0 \end{bmatrix} \phi(\xi) \qquad (a = \lfloor \xi \rfloor).$$

For a positive integer we write Φ^l for the l-fold composition of ϕ and we let $\phi^0(x) = x$.

Lemma. *Assume that n is not a perfect square. The sequence obtained by applying ϕ successively to $\sqrt{n}$ becomes periodic. More precisely, there exists an integer $l \geq 1$ such that*

$$\sqrt{n}, \quad \phi(\sqrt{n}), \quad \phi^2(\sqrt{n}), \quad \ldots, \quad \phi^{l+1}(\sqrt{n}) = \phi(\sqrt{n}),$$

Proof. We give here an ad-hoc proof customized to our situation. (For a more conceptual explanation see Section **??**). We claim that $\phi^k(\sqrt{n})$ is of the form $(b + \sqrt{n})/a$ with integers b and a such that $a|(b^2 - n)$. Indeed, this is true for $l = 0$, and if $\phi^k(\sqrt{n})$ is of the given form then, setting $s := \lfloor (b + \sqrt{n})/a \rfloor$, we have

$$\phi^k(\sqrt{n}) = \frac{1}{(b + \sqrt{n})/a - s} = a\frac{as - b + \sqrt{n}}{n - (as - b)^2},$$

which is again of the claimed form. Moreover, for $k \geq 1$, we have $\phi^k(\sqrt{n}) > 1$ and $0 > \phi^k(\sqrt{n})' > -1$, where, for rational number u and v, we use $(u + v\sqrt{n})' = u - v\sqrt{n}$. This can again be shown easily by induction. Finally, if we let X be the set of all numbers

$\xi := (b+\sqrt{n})/a$ with integers a and b such that $a|(n-b^2)$ and $\xi > 1$ and $0 > \xi' > -1$, then ϕ is injective on X as the reader can verify. But X is finite: the two inequalities imply $\xi\xi' < 0$, i.e. $b^2 < n$, and the divisibility condition bounds a. Thus ϕ permutes the elements of X and the lemma becomes now clear. □

Theorem. *Assume n is not a perfect square. Then the equation $x^2 - ny^2 = 1$ possesses a positive integral solution.*

Proof. With the notations as in the lemma we have

$$\sqrt{n} = \begin{bmatrix} a_0 & 1 \\ 1 & 0 \end{bmatrix} \phi(\sqrt{n}) = \begin{bmatrix} a_0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a_1 & 1 \\ 1 & 0 \end{bmatrix} \phi^2(\sqrt{n}) = \dots$$
$$\dots = \begin{bmatrix} a_0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a_1 & 1 \\ 1 & 0 \end{bmatrix} \cdots \begin{bmatrix} a_l & 1 \\ 1 & 0 \end{bmatrix} \phi^{l+1}(\sqrt{n}),$$

which, using $\phi^{l+1}(\sqrt{n}) = \phi(\sqrt{n})$ and $\phi(\sqrt{n}) = \left[\begin{smallmatrix} 0 & 1 \\ 1 & -a_0 \end{smallmatrix}\right] \sqrt{n}$, yields

$$\sqrt{n} = \begin{bmatrix} a_0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a_1 & 1 \\ 1 & 0 \end{bmatrix} \cdots \begin{bmatrix} a_l & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a_0 & 1 \\ 1 & 0 \end{bmatrix}^{-1} \sqrt{n}.$$

The matrix A occurring here has obviously integral entries. Since $a_1, \dots, a_l \geq 1$ the matrix is different from $\pm\left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$. The determinant of a matrix of the form $\left[\begin{smallmatrix} a & 1 \\ 1 & 0 \end{smallmatrix}\right]$ is -1, and hence the determinant of A equals $(-1)^l$. Applying again ϕ l-many times if necessary we can assume that l is even, so that A has determinant $+1$ and is hence in S_n. But then A represents a solution of $x^2 - ny^2 = 1$. Note that this solution is positive since it equals the first column of the matrix $\left[\begin{smallmatrix} a_0 & 1 \\ 1 & 0 \end{smallmatrix}\right] \cdots \left[\begin{smallmatrix} a_l & 1 \\ 1 & 0 \end{smallmatrix}\right]$. □

Even more is true.

Theorem. *Let l be the smallest positive even integer for which one has $\phi^{l+1}(\sqrt{n}) = \phi(\sqrt{n})$, and let $a_k := \lfloor \phi^k(\sqrt{n}) \rfloor$. Then*

$$F := \begin{bmatrix} a_0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a_1 & 1 \\ 1 & 0 \end{bmatrix} \cdots \begin{bmatrix} a_l & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a_0 & 1 \\ 1 & 0 \end{bmatrix}^{-1}$$

is the matrix representing the fundamental solution of $x^2 - ny^2 = 1$.

n	x	y	l	n	x	y	l
2	3	2	2	27	26	5	2
3	2	1	2	28	127	24	4
5	9	4	2	29	9801	1820	10
6	5	2	2	30	11	2	2
7	8	3	4	31	1520	273	8
8	3	1	2	32	17	3	4
10	19	6	2	33	23	4	4
11	10	3	2	34	35	6	4
12	7	2	2	35	6	1	2
13	649	180	10	37	73	12	2
14	15	4	4	38	37	6	2
15	4	1	2	39	25	4	2
17	33	8	2	40	19	3	2
18	17	4	2	41	2049	320	6
19	170	39	6	42	13	2	2
20	9	2	2	43	3482	531	10
21	55	12	6	44	199	30	8
22	197	42	6	45	161	24	6
23	24	5	4	46	24335	3588	12
24	5	1	2	47	48	7	4
26	51	10	2	48	7	1	2

Table 2. Fundamental solutions of equations $x^2 - ny^2 = 1$ for all $n < 50$ not a perfect square with period length l as in the theorem.

Proof. We have already shown than an l as in the theorem exists and that the corresponding M yields a solution to $x^2 - ny^2 = 1$. It remains to show that, for every positive solution there is an (even) $l \geq$ with $\phi^{l+1}(\sqrt{n}) = \phi^l(\sqrt{n})$ such that the matrix A derived from this l as above corresponds to the given solution. We postpone the main step in the proof of this to Section **??**. However, we can indicate at this point where this fact comes from.

For this let x, y be a positive solution of $x^2 - ny^2 = 1$. A quick calculation shows that

$$0 < \frac{x}{y} - \sqrt{n} = \frac{1}{(x + y\sqrt{n})y} < \frac{1}{2y^2},$$

where the inequality follows using that $x^2 = 1 + ny^2 > y^2$, so that $x + y\sqrt{n} > 2y$. Indeed if we look at the example $n = \sqrt{2}$ we see that $\frac{99}{70} - \sqrt{2} < \frac{1}{4900}$, and $\frac{90}{70} = 1.41428\ldots$, $\sqrt{2} = 1.41421\ldots$. As we shall see in Section **??** the above inequality for x/y means that the positive solutions x, y provide excellent approximations to $\sqrt{n}$. In fact, in Section **??** we shall prove much more. Namely, the inequality $|\frac{x}{y} - \xi| < \frac{1}{2y^2}$, for a given positive real number ξ, is equivalent to the fact that x/y is a convergent in the continued fraction expansion of ξ, which, as explained in the theory of continued fractions, means that the matrix of x, y equals one of the As as constructed above. □

The last theorem is quickly turned into an algorithm which can then be used to produce tables like Table 2.

Algorithm: Computation of the fundamental solution of Pell's equation $x^2 - ny^2 = 1$

```
def Pell_fs( n):
    """
    Returns the fundamental solution of
    x^2-ny^2=1.
    """
    U = lambda a: matrix( ZZ, 2, [a,1,1,0])
    phi = lambda x: 1/(x - floor(x))
    w = QQbar(sqrt(n))
    s = phi(w); lst = [w,s]
    k = 2; t = phi(s)
    while is_even(k) or t != s:
        lst.append(t)
        k += 1; t = phi(t)
    cf = map( lambda t: floor(t), lst)
```

```
    M = prod( U(a) for a in cf)*U(cf[0])^-1
    return M[0,0],M[1,0],len(cf)-1
```

Example. We consider the equation $x^2 - 23y^2 = 1$. Here we find:

k	0	1	2	3	4	5
$\phi^k(\sqrt{23})$	$\sqrt{23}$	$\frac{\sqrt{23}+4}{7}$	$\frac{\sqrt{23}+3}{2}$	$\frac{\sqrt{23}+3}{7}$	$\sqrt{23}+4$	$\frac{\sqrt{23}+4}{7}$
a_k	4	1	3	1	8	–

We compute

$$F = \begin{bmatrix}4 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}3 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}8 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}4 & 1\\ 1 & 0\end{bmatrix}^{-1} = \begin{bmatrix}24 & 115\\ 5 & 24\end{bmatrix}.$$

The fundamental solution is therefore $(24, 5)$.

When setting up a table as in the previous example the following hint might be helpful. It is easy to prove by induction that $\phi^k(\sqrt{n})$ is always of the form $\xi = \frac{\sqrt{n}+s}{t}$ with integers s and $t \geq 1$, and such that t divides $n - s^2$ (see the proof of the above lemma). The key is to write

$$\phi(\xi) = \frac{1}{(\sqrt{n}+s)/t - a} = \frac{\sqrt{n}+s+ta}{(n-(s-ta)^2)/t},$$

where a is the floor of ξ.

Example. The fundamental solutions of Pell's equation can become rather large. The fundamental solution of $x^2 - 1021y^2 = 1$ has for example fundamental solution

$$x = 1987238676909775732196682522310774156363518018 01$$
$$y = 6219237759214762827187409503019432615976684540.$$

15. Legendre's theorem

We consider a diophantine equation of the form

$$ax^2 + by^2 + cz^2 = 0,$$

where $abc \neq 0$, and where we pose the problem whether there exists a non-trivial solution in integers or, equivalently, in rational integers. The *trivial* solution is $x = y = z = 0$. Before stating a complete

answer we do some reductions. Obviously we can assume that the integers a, b, c are square-free (if, for example, $a = a_0 a_1^2$ with square-free a_0, we can substitute $x/a_1 \mapsto x$). We can further assume that a, b, c are relatively prime (by dividing the equation through the gcd of a, b, c if necessary). We can even assume that the three coefficients are pairwise relatively prime.

For seeing this assume that a and b have t as gcd. We multiply our equation by t and make the substitution $x/t \mapsto x$, $y/t \mapsto y$. Applying this procedure also to b, c and c, a we obtain an equation with pairwise relatively prime factors. Doing these reduction steps we arrive at an equation such that abc is squarefree.

Theorem (Legendre)**.** *Let a, b, c be integers such that abc is square-free. The equation $ax^2+by^2+cz^2 = 0$ possesses a non-trivial solution in integers if and only if the following conditions are satisfied:*

(1) *There exist a non-trivial solution in real numbers, and*

(2) *One has $\left(\frac{-ab}{p}\right) = \left(\frac{-bc}{q}\right) = \left(\frac{-ca}{r}\right) = 1$ for all odd primes $p \mid c$, $q \mid a$, $r \mid b$.*

Remark. Note that condition (1) can be quickly checked: we can find a solutions in real numbers if and only if a, b, c do not all have the same sign.

As we shall see in the proof the condition (2), for a prime $p \mid c$, is equivalent to the existence of a solution of $ax^2 + by^2 \equiv 0 \bmod p$ with x and y not both divisible by p.