Introduction to number theory

Fall 2019

Eero Saksman[1]

Department of mathematics and statistics
University of Helsinki

[1] I want to express my sincere gratitude for Sara Pyyköla and Paul Romanowski for their excellent LaTexification of an earlier handwritten edition of the notes, and many nice suggestions for new exercises!

Contents

"Mathematics is the queen of the sciences and number theory is the queen of mathematics." — Carl Friedrich Gauss

Chapter 1

Divisibility, primes and the greatest common divisor

1.1 Divisibility and primes

[1]

Definition 1.1. Let $a, b \in \mathbb{Z}$. Then a *divides* b, or, equivalently, a is *a factor* of b, if

$$b = k \cdot a, \text{where } k \in \mathbb{Z}$$

This property is marked with ” $\mid$ ”, e.g. $5 \mid 10$, $8 \mid 24$, and $-2 \mid 0$. The negation of this property, where a does not divide b or a is not a factor b, is marked with ” $\nmid$ ”, e.g. $4 \nmid 6$ and $15 \nmid 2$.

Agreement. From now on, $a, b, c, \ldots, x, y, z, \ldots$ are integers unless otherwise stated.

Theorem 1.2. *i) If $c \mid b$ and $b \mid a$, then $c \mid a$.*

ii) If $a \mid b$, then $a \mid bc \qquad \forall c$.

iii) If $a \mid b_1, a \mid b_2, \ldots, a \mid b_n$, then $a \mid (b_1 + b_2 + \cdots + b_n)$.

Proof. Left as an Exercise for the reader . □

Definition 1.3. If $p \in \mathbb{N}$, $p \geq 2$ and $k \in \mathbb{N}$, $k \mid p$, $\Rightarrow$ $k \in \{1, p\}$, then p is a *prime*. This is denoted as $p \in \mathbb{P}$. Thus $\mathbb{P} = \{2, 3, 5, 7, 11, \ldots\}$.

Definition 1.4. Let $n \in \mathbb{N}, n \geq 2$. If $n \notin \mathbb{P}$, then n is a *composite* number.

[1] I want to express my sincere gratitude for Sara Pyyköla and Paul Romanowski for their excellent LaTexification of an earlier handwritten edition of the notes, and many nice suggestions for new exercises!

Theorem 1.5. *Every natural number $n \geq 2$ is a product of primes.*

Proof. Assume the contrary: there is at least one positive integer ≥ 2 which is not a product of primes. Let n be the smallest such number; then, $n \notin \mathbb{P}$, whence $n_1 \cdot n_2 = n$, where $n_1, n_2 > 1$. This implies also that $n_1, n_2 < n$. Now n_1 and n_2 are products of primes, so n is also, which is a contradiction. □

Theorem 1.6. *There are infinitely many primes.*

Proof. Assume the contrary: $\mathbb{P}$ is finite, so $\mathbb{P} = \{p_1, p_2, p_3, \ldots, p_k\}$. Then, if $n := 1 + p_1 \cdot p_2 \cdot \ldots \cdot p_k$, we have $p_j \nmid n$ for $\forall j = 1, 2, \ldots, k$, which contradicts Theorem 1.5. Therefore $\mathbb{P}$ is infinite. □

The previous result was already given in Euklid's famous book series on geometry 'Elements' (around 300BC)!

1.2 Greatest common divisor and uniqueness of prime number decomposition

Theorem 1.7. (*The remainder theorem*) *Let $b \geq 1$. Then for any integer a there are unique k, r, such that*

$$a = kb + r, \;\; r \in \{0, 1, \ldots, b - 1\}.$$

Proof. Let's prove the existence of k and r first. Denote $U := \{a - ub \mid a - ub \geq 0 \text{ and } u \in \mathbb{Z}\}$. Clearly $U \neq \varnothing$. Let r be the smallest element in U, and write $r = a - u_1 b$. If $r \geq b$, we have

$$r > r - b = a - (u_1 + 1)b \geq 0,$$

which is a contradiction with the definition of r. Thus $0 \leq r < b$, and we may choose $k = u_1$.

Now let's prove that k and r are unique. If $0 \leq r_1, r_2 \leq b$ and

$$r_1 = a - u_1 b \quad \text{and} \quad r_2 = a - u_2 b,$$

then $b \mid (r_1 - r_2)$. Now $|r_1 - r_2| < b$, whence $r_1 = r_2$. □

The following result is fundamental, and its truthfulness is more nontrivial than one might guess by first thought.

Theorem 1.8. *Let $a \neq 0$ and $b \neq 0$. There is a unique integer d with the properties*

i) $d \,|\, a$ *and* $d \,|\, b$

ii) *if* $d' \,|\, a$ *and* $d' \,|\, b \Rightarrow d' \,|\, d$

iii) $d \geq 1$.

d is called the greatest common divisor *of a and b and denoted by $gcd(a, b)$.*

Proof. Let us set

$$d = \min\{ax + by \mid ax + by \geq 1,\; x, y \in \mathbb{Z}\} \tag{1.1}$$

Clearly the set on the right-hand side is non-empty , so d is well-defined. By the definition of d it satisfies (iii), and we may write for some x_0, y_0

$$d = ax_0 + by_0.$$

This implies that d fulfils $ii)$. In order to check $i)$, assume to the contrary that $d \nmid a$. Then by the Theorem 1.7, we have

$$a = dk + r, \quad 0 < r < d,$$

from which we can obtain

$$r = a - dk = a - (ax_0 + by_0)k = (1 - kx_0)a + (-ky_0)b.$$

This contradicts the definition of d since $1 \leq r < d$. Thus $d \,|\, a$ and a similar reasoning shows that $d \,|\, b$. Uniqueness follows by noting that if both d_1 and d_2 satisfy $i) - iii)$, then especially $d_1 \,|\, d_2$ and $d_2 \,|\, d_1$. This implies that $d_1 = d_2$ since the numbers are positive. □

Definition 1.9. If $gcd(a, b) = 1$, we say that a and b are *relative primes* or *co-primes.*

Corollary 1.10. Let $a \neq 0, b \neq 0$ and $d = gcd(a, b)$. Then

i) $d = ax_0 + by_0$ for some $x_0, y_0 \in \mathbb{Z}$

ii) $\{xa + by \mid x, y \in \mathbb{Z}\} = \{kd \mid k \in \mathbb{Z}\}$

Proof. $i)$ follows from (1.1) in the proof of Theorem 1.8. Relation LHS $\supset$ RHS in $ii)$ follows from $i)$. In turn, LHS $\subset$ RHS in $ii)$ follows by noting that $d \,|\, a$ and $d \,|\, b$, which shows that $d \,|\, (ax + by)$ for any x, y. □

Example. $gcd(8, 22) = 2$, and we may write $2 = 3 \cdot 22 - 8 \cdot 8$.

Theorem 1.11. *i) If $d > 1$, then $gcd(a,b) = d \Leftrightarrow gcd\left(\frac{a}{d}, \frac{b}{d}\right) = 1$.*

ii) $gcd(a,b) = gcd((a+kb), b)$.

Proof. Left as an Exercise **??** for the reader . □

Theorem 1.12. *If $a \,|\, bc$ and $gcd(a,b) = 1$, then $a \,|\, c$.*

Proof. By Corollary 1.10 we have $ax_0 + by_0 = 1$ for some x_0, y_0. Then $c = acx_0 + bcy_0$, where $a \,|\, acx_0$ and $a \,|\, bcy_0$. □

Corollary 1.13. If $p \in \mathbb{P}$ and $p \,|\, ab$, then $p \,|\, a$ or $p \,|\, b$.

Proof. Follows from Theorem 1.12. □

Note that by induction we deduce that if $p \in \mathbb{P}$ and $p \,|\, a_1 a_2 \ldots a_k$, then $p \,|\, a_j$ for some $j \in \{1, 2, 3, \ldots, k\}$.

Corollary 1.14. If $p \,|\, q_1 q_2 \ldots q_k$, where $p, q_1, q_2, q_3, \ldots, q_k \in \mathbb{P}$, then $p = q_j$ for some j.

Proof. We must have $p \,|\, q_j$ for some j, but then $p = q_j$, since $p, q_j \in \mathbb{P}$. □

We are now ready for the "Fundamental Theorem of Arithmetics". Here the key new feature is the uniqueness statement.

Theorem 1.15. *Every positive integer $n \geq 2$ can be written in the form*

$$n = p_1 p_2 p_3 \ldots p_k,$$

where $p_1, p_2, p_3, \ldots, p_r \in \mathbb{P}$. This representation is unique up to the order of the factors.

Proof. By Theorem 1.5 every $n \geq 2$ is a product of primes. Assume to the contrary that the claim is not true. Then we may consider $n \geq 2$ which is the smallest positive integer with two representations

$$n = p_1 p_2 p_3 \ldots p_k = q_1 q_2 q_3 \ldots q_m,$$

(here $p_1, p_2, \ldots, p_k, q_1, q_2, \ldots, q_m \in \mathbb{P}$) that cannot be made to coincide even by reordering. Now $p_1 \,|\, q_1 q_2 \ldots q_m$, whence by Corollary 1.14 $p_1 = g_j$ for some $1 \leq j \leq m$. Dividing by p_1

we see that the positive integer $\dfrac{n}{p_1} < n$ has also two essentially different representations which contradicts the definition of n. □

By Theorem 1.15 every integer $n \geq 1$ can be uniquely written as

$$n = \prod_{k=1}^{\infty} p_k^{\alpha_k}$$

where $\alpha_k \geq 0$ and $\alpha_n > 0$ only for finitely many k. In addition, $p_1 < p_2 < p_3 < \dots$ are the primes in increasing order. Thus

$$p_1 = 2,\, p_2 = 3,\, p_3 = 5, \dots$$

If $n = \prod_{k=1}^{\infty} p_k^{\alpha_k}$ and $m = \prod_{k=1}^{\infty} p_k^{\beta_k}$, then

$$gcd(n,m) = \prod_{k=1}^{\infty} p_k^{\gamma_k}, \qquad \gamma_k = \min\{\alpha_k, \beta_k\},\, k \geq 1.$$

Exercise. Generalize Theorem 1.8 into the following Definition 1.16.

Definition 1.16. Let $a_1, a_2, \dots, a_k$ be integers, which are not all zero. Then there exists a unique integer d, denoted by $d = gcd(a_1, a_2, \dots, a_n)$ with the properties

i) $d \,|\, a_k, 1 \leq k \leq n$,

ii) if $d' \,|\, a_k, 1 \leq k \leq n$, then $d' = d$

iii) $d \geq 1$

Moreover, $d = \sum_{j=1}^{n} x_j a_j$ holds for some integers $x_1, x_2, x_3, \dots, x_n$.

Definition 1.17. Let $a, b \geq 1$. Then there exists a unique integer h, such that

i) $a \,|\, h$ and $b \,|\, h$

ii) if $a \,|\, h'$ and $b \,|\, h' \Rightarrow h \,|\, h'$

iii) $h \geq 1$.

The integer h is called *the least common multiple* of a and b and is denoted by $lcm(a, b)$.

Proof. Left as an Exercise for the reader. □

Example 1. lcm(12, 15) = 60.

1.3 Linear Diophantine equations and the Euclidean algorithm

Theorem 1.18. *Let $a \neq 0$ and $b \neq 0$. The equation $ax + by = c$ has a solution in integers if and only if $d \,|\, c$, where $d = gcd(a, b)$. All the solutions are obtained from the formula*

$$\begin{cases} x = x_1 + \dfrac{bt}{d}, \\ y = y_1 - \dfrac{at}{d}, t \in \mathbb{Z} \end{cases}$$

where $\{x_1, y_1\}$ is one solution.

Proof. Write $d = gcd(a, b)$. The sufficiency and necessity of the given condition is clear from the condition *ii*) of Corollary 1.10 ! Assume then that $d \,|\, \{a, b\}$. We may divide the equation $ax + by = c$ by d to obtain equivalently

$$a'x + b'y = c'$$

where $a' = \dfrac{a}{d}$, $b' = \dfrac{b}{d}$ and $c' = \dfrac{c}{d}$. Then $gcd(a', b') = 1$ by Theorem 1.12. If $\{x_1, y_1\}$ and $\{x_2, y_2\}$ are solutions, we obtain

$$a'x_1 + b'y_1 = a'x_2 + b'y_2, \quad \text{or} \quad a'(x_2 - x_1) = b'(y_1 - y_2).$$

Since $gcd(a', b') = 1$ and $b' \,|\, \{a'(x_2 - x_1)\}$, we must have $b' \,|\, (x_2 - x_1)$, or

$$x_2 = x_1 + tb, \qquad \text{for some} \quad t \in \mathbb{Z}.$$

Then $y_2 = y_1 - ta$. Conversely, direct substitution (do it!) shows that all pairs given by the formula solve the equation. □

While the previous theorem is beautiful, how are we to find some solution to the equation $ax + by = c$ without burdensome trial and error? A solution to this problem is given by the following algorithm.
First, let's assume that $a > b > 1$. To find $gcd(a, b)$, we'll use the remainder theorem (Theorem 1.7) to write consecutively

$$\begin{aligned} a &= bq_1 + r_1, & 0 < r_1 < b \\ b &= r_1 q_2 + r_2, & 0 < r_2 < r_1 \\ r_1 &= r_2 q_3 + r_3, & 0 < r_3 < r_2 \\ &\vdots & \vdots \\ r_{k-2} &= r_{k-1} q_k + r_k, & 0 < r_k < r_{k-1} \\ r_{k-1} &= r_k q_{k+1}, \end{aligned}$$

where the remainder is zero for the first time in the last equation.

Theorem 1.19. $r_k = gcd(a, b)$.

Proof. From the equations in the Euclidean algorithm we can obtain

$$\begin{aligned}
\text{Last equation} &\Rightarrow \quad r_k \,|\, r_{k-1}\\
\text{Second last equation} &\Rightarrow \quad r_k \,|\, r_{k-2}\\
&\vdots\\
\text{Second equation} &\Rightarrow \quad r_k \,|\, b\\
\text{First equation} &\Rightarrow \quad r_k \,|\, a
\end{aligned}$$

Since $r_k \,|\, b$ and $r_k \,|\, a$ we deduce that $r_k \,|\, d$, where $d = gcd(a, b)$. Conversely, it also holds that

$$\begin{aligned}
\text{First equation} &\Rightarrow \quad d \,|\, r_1\\
\text{Second equation} &\Rightarrow \quad d \,|\, r_2\\
&\vdots\\
\text{Second last equation} &\Rightarrow \quad d \,|\, r_k.
\end{aligned}$$

Thus both $d \,|\, r_k$ and $r_k \,|\, d$. Altogether, since $r_k \geq 1$, we must have $r_k = d$. □

The Euclidean algorithm may also be used to solve the equation $ax + by = d$, where $d = gcd(a, b)$, in a following fashion:

$$\begin{aligned}
\text{First equation} &\Rightarrow \quad r_1 = \; a - bq_1 = u_1 a + u_1 b\\
\text{Second equation} &\Rightarrow \quad r_2 = \; b - q_2 r_1 = b - q_2(u_1 a + u_1 b) = u_2 a + u_2 b\\
&\vdots\\
\text{Second last equation} &\Rightarrow \quad d = r_k = r_{k-2} - q_k r_{k-1}\\
&\qquad = (u_{k-2} a + u_{k-2} b) - q_k(u_{k-1} a + u_{k-1} b)\\
&\qquad = a(u_{k-2} - q_k u_{k-1}) + b(u_{k-2} - q_k u_{k-1})\\
&\qquad = ax_0 + by_0
\end{aligned}$$

Naturally in practise one doesn't need to use the unwieldy notation above.

Example. $127x - 87y = 1$ is solvable since $gcd(127, 87) = 1$. Indeed, by the Euclidean algorithm we have

$$\begin{aligned}
127 &= 87 \cdot 1 + 40\\
87 &= 40 \cdot 2 + 7\\
40 &= 7 \cdot 5 + 5\\
7 &= 5 \cdot 1 + 2\\
5 &= 2 \cdot 2 + 1\\
(2 &= 1 \cdot 2 + 0)
\end{aligned}$$

Denote $127 = a$ and $87 = b$. We obtain successively

$$\begin{aligned}
40 &= a - b\\
7 &= b - 2(a - b) = -2a + 3b\\
5 &= (a - b) - 5(-2a + 3b) = 11a - 16b\\
2 &= (-2a + 3b) - (11a - 16b) = -13a + 17b\\
1 &= (11a - 16b) - 2(-13a + 19b) = 37a - 54b
\end{aligned}$$

Thus $\{x_0, y_0\} = \{37, 54\}$ is one solution and the general solution is $\{x_0, y_0\} = \{37 + 87t,\ 54 + 127t \mid t \in \mathbb{Z}\}$.

The Euclidean algorithm is actually rather effective for large numbers, as the following theorem demonstrates.

Theorem 1.20. *Let $1 < a < b$. The number of steps needed to compute gcd(a, b) by the Euclidean algorithm does not exceed $5 \log_{10} a$.*

Proof. We leave this as an exercise for interested readers. □

Chapter 2

Congruences and ring $\mathbb{Z}_m$

2.1 Congruences

Great mathematician Carl Friedrich Gauss initiated the use of ”congruence”[1]. It simplifies many considerations about divisibility.

Definition 2.1. Let $m \neq 0$. If $m \mid (a-b)$, we say that a is *congruent* to b modulo m, which is denoted by

$$a \equiv b \pmod{m}$$

Note that often when it is absolutely clear from the context, one may leave out the modulus m.

Note that plural of modulus is moduli. Thus, we might write $a \equiv b \pmod{m, n}$, which means that both m and n divide $(a-b)$, Then the moduli are m and n.
The following Lemma 2.2 verifies that congruence is a honest equivalence relation!

Lemma 2.2. The congruence relation has the following properties (which characterise an equivalence relation):

1. $a \equiv a \pmod{m}$ (reflexivity)
2. $a \equiv b \pmod{m} \Leftrightarrow b \equiv a \pmod{m}$ (symmetry)
3. $a \equiv b \pmod{m}$ and $b \equiv c \pmod{m} \Rightarrow a \equiv c \pmod{m}$ (transitivity)

Proof. Follows directly from the Definition 2.1. □

The following result verifies that when valid congruences are summed or multiplied sidewise and one gets valid congruences.

[1] The notion of congruence is one of the first examples in the whole history of mathematics of usefulness of abstract notions in dealing with concrete problems.

Theorem 2.3. *i)* $a \equiv b \pmod{m}$ *and* $k \in \mathbb{Z}$*, then*

$$a + k \equiv b + k \pmod{m} \quad \textit{and} \quad ka \equiv kb \pmod{m}.$$

ii) $a \equiv b \pmod{m}$ *and* $c \equiv d \pmod{m}$*, then*

$$a + c \equiv b + d \pmod{m} \quad \textit{and} \quad ab \equiv cd \pmod{m}.$$

Proof. Left as an Exercise **??** for the reader. □

Theorem 2.4. *If* $P(x)$ *is a polynomial with integer coefficients, then* $a \equiv b \pmod{m} \Rightarrow P(a) \equiv P(b) \pmod{m}$.

Proof. If $a \equiv b \pmod{\mathrm{m}}$, we may iterate Theorem (2.3) *ii)* and obtain $a^k \equiv b^k \pmod{\mathrm{m}}$, for $\forall k \geq 0$. If

$$P(x) = \sum_{k=0}^{l} c_k x^l,$$

we can especially obtain $c_k a^k \equiv c_k b^k \pmod{\mathrm{m}}$, from where the claim follows by summing up over $k = 0, 1, 2, \ldots, l$ (now we are iterating Theorem (2.3) *i)*). □

A similar reasoning yields the following theorem.

Theorem 2.5. *If* P *is a polynomial of* k *variables with integer coefficients, then*

$$P(x_1, x_2, \ldots, x_k) \equiv P(y_1, y_2, \ldots, y_k) \pmod{m}$$

whenever $x_j \equiv y_j \pmod{m}$, $j = 1, 2, \ldots, k$.

Definition 2.6. If $a \equiv b \pmod{m}$ does not hold, we denote $a \not\equiv b \pmod{m}$.

Example. $7 \equiv -13 \pmod{20}$, $7 \equiv 67 \pmod{20}$, $200 \equiv -10 \pmod{3}$.

Example. Show that the number $n = 59(168^7 + 87^7) + 9$ is divisible by 13.

Solution. $59 \equiv 7 \pmod{13}$, $168 \equiv -1 \pmod{13}$ and $87 \equiv -4 \pmod{13}$. Thus

$$n \equiv 7((-1)^7 + (-4)^7) + 9 \pmod{13}.$$

Now $(-4)^2 \equiv 3 \pmod{13}, (-4)^4 \equiv 3^2 = 9 \pmod{13}, (-4)^6 \equiv 9 \cdot 3 \equiv 1 \pmod{13}$ and $(-4)^7 \equiv -4 \cdot 1 = -4 \pmod{13}$. Finally

$$n \equiv 7(-1 - 4) + 9 \equiv -26 \equiv 0 \pmod{13}.$$

Remark. When computing $a^k \bmod m$, it is sometimes useful to complete inductively $a^1, a^2, a^4, a^8, \ldots$ modulo m, and express a^k using them, e.g.:

$$a^{23} = a^{16} a^4 a^2 a^1$$

Example. We have $10 \equiv 1 \pmod 9$. Hence, if $S(n)$ stands for the number of digits of n in base 10, we have

$$n \equiv S(n) \equiv S(S(n)) \equiv \ldots (\text{mod } 9)$$

For demonstration, let $n = 5476289$. Then $S(n) = 5+4+7+6+2+8+9 = 41$, $S(S(n)) = 4+1 = 5 \not\equiv 0 \pmod 9$, thus $9 \nmid n$.

Theorem 2.7. *If $ac \equiv bc \pmod m$ and $d = gcd(m,c)$, then $a \equiv b \pmod{\frac{m}{d}}$.*

Proof. By assumption $c(a-b) = km$ for some k. If we divide this by d it follows that $c'(a-b) = km'$, where $c' = \frac{c}{d}$ and $m' = \frac{m}{d}$. From this can see that $m' \mid c'(a-b)$ and by the 1.12, $gcd(m', c') = 1$ so $m' \mid (a-b)$, which is the claim. □

Problem. One moves a clock 23 minutes ahead once a day. After how many days does it show the time 23.59? The starting time is 00.00.

Solution. Equivalently, we search for the smallest positive integer $x \geq 1$ so that it satisfies the congruence

$$23x \equiv -1 \pmod{24 \cdot 60}$$

which is a Diophantine equation of the following form:

$$23x + 1440y = -1$$

and we look for pair (x, y) with smallest positive x. We apply the Euclidean algorithm we find that the smallest x is 313, since $gcd(23, 1440) = 1$: (we denote $a := 1440$ and $b := 23$)

$$\begin{aligned}
1440 &= 62 \cdot 23 + 14\\
23 &= 1 \cdot 14 + 9\\
14 &= 1 \cdot 9 + 5\\
9 &= 1 \cdot 5 + 4\\
5 &= 1 \cdot 4 + 1\\
4 &= 4 \cdot 1 \Rightarrow\\
14 &= 1440 - 62 \cdot 23 = a - 62b\\
9 &= 23 - 14 = 63b - a\\
5 &= 14 - 9 = 2a - 125b\\
4 &= 9 - 5 = 188b - 3a\\
1 &= 5 - 4 = -313b + 5a.
\end{aligned}$$

Especially, $23x_0 + 1440y_0 = -1$, where $(x_0, y_0) = (313, -5)$. According to Theorem 1.18 the general solution for x-values is $x = 313 + k1440$, $k \in \mathbb{Z}$, and clearly the smallest positive x is given by $x_0 = 313$.

At this point, we'll need to recall a few concepts from algebra before we can go further with the congruences.

2.2 Short review of basic algebra

Definition 2.8. $(S, \circ)$ is a *group* if the operation $\circ\colon S \times S \to S$, denoted $\circ(a,b) = a \circ b$, has the following properties:

i) $a \circ (b \circ c) = (a \circ b) \circ c,\ \forall a, b, c \in S$ (Associativity)

ii) $\exists e \in S$ so that $e \circ a = a \circ e = a,\ \forall a \in S$; e is called a *neutral element*

iii) Every $a \in S$ has an *inverse* $a^{-1} \in S$ so that $a \circ a^{-1} = a^{-1} \circ a = e$.

Remark. If only *i)* and *ii)* hold, we say that $(S,\ \circ)$ is a monoid.

Examples.

1. $(\mathbb{Z},\ +)$, $(\mathbb{Q},\ +)$, $(\mathbb{R},\ +)$ are groups, when $+$ is the standard addition.
2. $(\mathbb{Q} \setminus \{0\},\ \cdot)$ is a group, when $\cdot$ is the standard multiplication.
3. $(\{0,\ 1\},\ \oplus)$ is a group with two elements if define the ”$\oplus$” followingly:

$\oplus$	0	1
0	0	1
1	1	0

4. $(\mathbb{Q}, \cdot)$ is a monoid when $\cdot$ is the standard multiplication.

Definition 2.9. The operation $\circ$ is *commutative*, if $a \circ b = b \circ a,\ \forall a, b \in S$. Groups or monoids with commutative operations are called *Abelian* groups or monoids.

Definition 2.10. $(S, \oplus, \odot)$ is an Abelian *ring* if it has the following properties:

i) $(S, \oplus)$ is an Abelian group, with a neutral element denoted by e

ii) $(S, \odot)$ is an Abelian monoid, with a neutral element denoted by 1

iii) $e \neq 1$

iv) $a \odot (b \oplus c) = (a \odot b) \oplus (a \odot c),\quad \forall a, b, c \in S$ (Distributivity)

Definition 2.11. $(S, \oplus, \odot)$ is a *field*, if it is an Abelian ring and $(S \setminus \{e\}, \odot)$ is a group.

Remark. An Abelian ring is a field if and only if every non-zero element has a multiplicative inverse.

Examples.

i) $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$ are rings with the standard addition and multiplication. $\mathbb{Q}$ and $\mathbb{R}$ are fields.

ii) $(\{0,1\}, \oplus, \odot)$ is a field if we set

$\oplus$	0	1
0	0	1
1	1	0

$\odot$	0	1
0	0	0
1	0	1

The proof of this is left as an Exercise **??** (see page **??**).

For the purpose of this course, we need only these definitions, so let's get back to congruences.

2.3 The ring $\mathbb{Z}_m$ of residue classes

According to the Theorem 2.2, congruence is an equivalence relation. The corresponding equivalence classes are called *residue classes* or *congruence classes* modulo m.

Lemma 2.12. There are m different residue classes for modulus m.

Proof. If $n \in \mathbb{Z}$, by the Theorem 1.7 we may write

$$n = km + r, \quad 0 \geq r \geq m - 1.$$

Then $n \equiv r \pmod{m}$, so every number is congruent to one of the numbers $0, 1, \dots, m-1$. Conversely, all the numbers are clearly non-congruent modulo m, and their number is m. □

Definition 2.13. The residue class defined by the given number a and the modulus m is the set

$$\overline{a} := \{x \in \mathbb{Z} \mid x \equiv a \pmod{m}\} = \{a + km \mid k \in \mathbb{Z}\},$$

which is denoted by $\lfloor a \rfloor_m$.

Definition 2.14. $\mathbb{Z}_m = \{\overline{a} \mid a \in \mathbb{Z}\} = \{\overline{0}, \overline{1}, \dots, \overline{m-1}\}$.

Remark. Instead of $\mathbb{Z}_m$, a more exact notation would be $\mathbb{Z}\backslash m\mathbb{Z}$, but $\mathbb{Z}_m$ is more practical and common.

Theorem 2.15. $(\mathbb{Z}_m, +, \cdot)$ *is an Abelian ring with* m *elements when* $+$ *('addition') and* $\cdot$ *('multiplication') in* $\mathbb{Z}_m$ *are defined by the rules*

$$\overline{a} + \overline{b} = \overline{a+b}, \quad \overline{a} \cdot \overline{b} = \overline{a \cdot b}, a, b \in \mathbb{Z}$$

Proof. Let us first check that e.g. $\overline{a}\overline{b}$ does not depend on the used representations a, b. Thus, let $\overline{a'} = a$ and $\overline{b'} = b$. Then $a \equiv a' \pmod{m}$ and $b' \equiv b \pmod{m}$ whence

$$\overline{a'b'} \equiv \overline{ab} \pmod{m}$$

which shows that multiplication is well-defined. Similar reasoning works for the addition. The properties of the ring are now easily checked, mostly by putting bars over the knwon facts in $\mathbb{Z}$. For example, when 0 denotes the the neutral element for addition in $\mathbb{Z}$:

$$0 + a = a + 0 = a, \quad \forall a \in \mathbb{Z} \quad \Rightarrow \quad \overline{0} + \overline{a} = \overline{a} + \overline{0} = \overline{a}, \quad \forall\ \overline{a} \in \mathbb{Z}_m.$$

Thus $\overline{0}$ is the neutral element for for addition in $\mathbb{Z}_m$. Other conditions of the ring are verified likewise, e.g. the commutativity of the multiplication comes straight from the definition:

$$\overline{a}\overline{b} = \overline{ab} = \overline{ba} = \overline{b}\overline{a}.$$

□

Example. One notes that always e.g. $\overline{a} - \overline{b} = \overline{a-b}$ etc. Thus in $\mathbb{Z}_6$ one (not always the easiest way) to compute is

$$\overline{2} - \overline{3}(\overline{5} - \overline{7}) = \overline{2 - 3(5-7)} = \overline{8} = \overline{2}.$$

Definition 2.16. By Theorem 1.7 or by the proof of Lemma 2.12, any $n \in \mathbb{Z}$ is congruent to exactly one number $r \in \{0, 1, \ldots, m-1\}$. Then r is called *the smallest positive remainder of* n modulo m.

Theorem 2.17. *Let* $m \geq 2$. *The ring* $\mathbb{Z}_m$ *is a field if and only if* $m \in \mathbb{P}$.

Proof. Assume first that $m \notin \mathbb{P}$. Then there are integers $m_1, m_2 > 1$, so that $m = m_1 m_2$. Especially $1 < m_1 < m$ so that $\overline{m_1} \neq \overline{0}$ in $\mathbb{Z}_m$. If $\mathbb{Z}_m$ is a field, then $\exists b \in \mathbb{Z}$ with $\overline{m_1}\overline{b} = \overline{1}$ or

$$1 \equiv m_1 b \pmod{m}$$

Multiplying by m_2 we obtain

$$m_2 \equiv m_1 m_2 b = mb \equiv 0 \pmod{m}$$

However, $1 < m_2 < m$, so this yields a contradiction. Hence $\mathbb{Z}_m$ is not a field.
Assume next that $m \in \mathbb{P}$. Given $\overline{a} \neq \overline{0}$ in $\mathbb{Z}_m$ we have that $m \nmid a$ and since $m \in \mathbb{P}$ we deduce that $gcd(m, a) = 1$. By *i)* of the Corollary 1.13, there are $x_0, y_0 \in \mathbb{Z}$ such that

$$1 = x_0 m + y_0 a \quad \Rightarrow \quad \overline{1} = \overline{x_0}\overline{m} + \overline{y_0}\overline{a} = \overline{y_0}\overline{a} =$$

so that $\overline{a}$ has the inverse $\overline{y_0} \in \mathbb{Z}_m$. Hence $\mathbb{Z}_m$ is a field. □

Definition 2.18. $a_1, a_2, \dots, a_m$ is a *complete residue system* modulo m, if

$$\{\overline{a_1}, \overline{a_2}, \dots, \overline{a_m} = \mathbb{Z}_m\}$$

Clearly this means that $\{\overline{a_1}, \overline{a_2}, \dots, \overline{a_m}\}$ contains exactly one element from each residue class, from which we can formulate the following result:

Theorem 2.19. *Let $m \geq 1$ and $U \subset \mathbb{Z}$. Then U is a complete residue system modulo m if and only if at least two of the following hold:*

i) U contains m elements

ii) any two elements in U are non-congruent modulo m

iii) For every integer a there is $u \in U$ so that $a \equiv u \pmod{m}$.

Proof. Left as an Exercise for the reader. □

Corollary 2.20. Let $\{a_1, a_2, \dots, a_m\}$ be a complete residue system modulo m, $gcd(k, m) = 1$ and $b \in \mathbb{Z}$. Then

$$A := \{ka_1 + b, ka_2 + b, \dots, ka_m + b\}$$

is also a complete residue system modulo m.

Proof. Since there are m elements in A, it is enough to check that they are mutually non-congruent modulo m. If $ka_j + b \equiv ka_l + b \pmod{m}$, we have that $m \,|\, k(a_j - a_l)$. Since $gcd(m, k) = 1$, it follows that $m \,|\, (a_j - a_l)$, hence $j = l$ by definition. □

The following theorem, called ”Fermat’s Little Theorem” (FLT for short), belongs to the fundamental results in number theory.

Theorem 2.21. *If $p \in \mathbb{P}$ and $p \nmid a$, then $a^{p-1} \equiv 1 \pmod{p}$.*

Proof. $\{0, 1, \dots, p - 1\}$ is a complete residue system modulo p, and so is
also $\{0, a, 2a, \dots, (p - 1)a\}$ by the Corollary 2.20 since $gcd(a, p) = 1$. Especially the

numbers $\{1, 2, \dots, p-1\}$ are congruent to numbers $\{a, 2a, \dots, (p-1)a\}$ in some order. Thus

$$a \cdot 2a \cdot \ldots \cdot (p-1)a \equiv 1 \cdot 2 \cdot \ldots \cdot (p-1) \pmod p \quad \Leftrightarrow \quad (p-1)!a^{p-1} \equiv (p-1)! \pmod p$$

Since $gcd((p-1)!, p) = 1$, we may divide the equation by $(p-1)!$ in accordance to Theorem 1.8 and obtain $a^{p-1} \equiv 1 \pmod p$. □

An immediate reformulation of the Theorem 2.21 gives us the following corollary.

Corollary 2.22. If $p \in \mathbb{P}$, then $a^p \equiv a \pmod{\text{p}}$ for $\forall a \in \mathbb{Z}$.

Both Theorem 2.21 and Corollary 2.22 are abbreviated as "FLT" (not to be mixed with Fermat's last theorem!).

Example. $p \mid 2^p - 2$ for all $p \in \mathbb{P}$.

Problem. Let $p \in \mathbb{P}$, $p \neq 2, 5$. Show that p divides infinitely many numbers in the sequence

$$1, 11, 111, 1111, \dots.$$

Solution. If the decimal expansion $11111\dots1$ contains k ones, we have $11111\dots1 = 10^{k-1} + 10^{k-2} + \ldots + 1 = \dfrac{10^k - 1}{9}$. Then we have two cases:

i) $p \neq 3$, whence $p \geq 7$. Then it is enough to show that $10^k \equiv 1 \pmod p$ for infinitely many k. By FLT we have $10^{p-1} \equiv 1 \pmod p$, since $gcd(10, p) = 1$. Multiplying this congruence by itself ℓ times we thus get $10^{\ell(p-1)} \equiv 1 \pmod p$, $\forall \ell \geq 1$.

ii) $p = 3$. Since $10 \equiv 1 \pmod 3$ we have $10^k \equiv 1 \pmod 3$, $\forall k \geq 1$. Thus $3 \mid 11111\dots1$ whenever the number of ones is divisible by 3. □

2.4 Euler's totient function

Definition 2.23. Euler's totient function (or Euler's φ-function) $\varphi\colon \mathbb{N} \to \mathbb{N}$ is defined by setting $\varphi(1) = 1$ and for $n \geq 2$, $\varphi(\text{n})$ is the number of elements $a \in \{1, 2, \dots, m\}$ such that $gcd(a, m) = 1$. In other words, Euler's totient function φ counts the positive and relatively prime integers up to a given integer n.

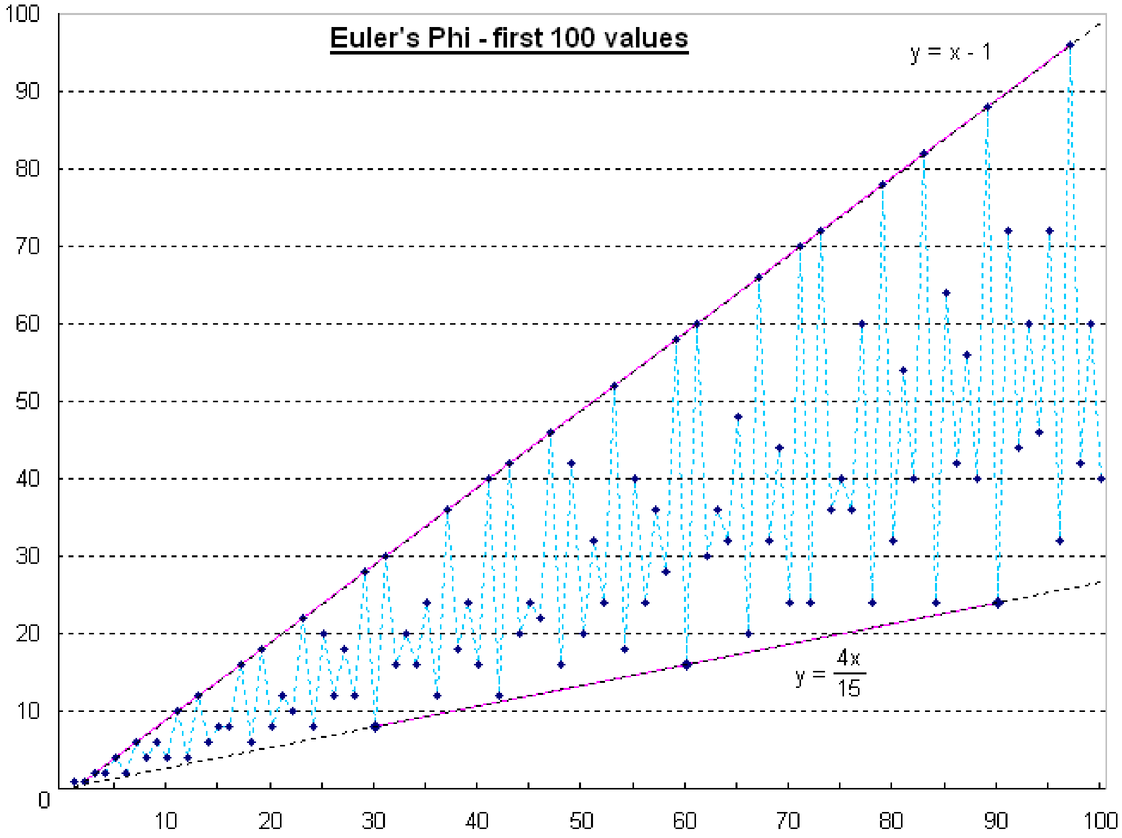

Figure 2.1: A figure about the values of $\varphi(n)$, $1 \leq n \leq 100$.

Source: https://commons.wikimedia.org/wiki/File:EulerPhi100.PNG. Retrieved 27.6.2019.

Note that the lower bound $y = \dfrac{4x}{15}$ in the Figure 2.1 only applies when $n \leq 100$!

Example. $\varphi(4) = 2$, $\varphi(5) = 4$, $\varphi(6) = 2$, $\varphi(7) = 6, \ldots$.

Definition 2.24. *A number theoretic function* is a function $f\colon \mathbb{N} \to \mathbb{R}$.

Definition 2.25. A number theoretic function f is multiplicative if $f(mn) = f(m)f(n)$ when $gcd(m,n) = 1$.

Theorem 2.26. *φ is multiplicative.*

Proof. Let us consider the array

$$\begin{array}{ccccc}
0 & 1 & 2 & \ldots & m-1 \\
m & m+1 & m+2 & \ldots & 2m-1 \\
\vdots & \vdots & \vdots & & \vdots \\
(n-1)m & (n-1)(m+1) & (n-1)(m+2) & \ldots & nm-1
\end{array}$$

Every column consist of numbers that are mutually congruent modulo m. By looking at the first row we deduce that there are $\varphi(m)$ columns whose elements are relatively prime with m. Any element from any other column has a non-trivial common factor with m.

On any fixed column there are exactly $\varphi(n)$ elements that are relatively prime to n since by the Corollary 2.20 any column forms a complete residue system modulo n.

Put together $\varphi(mn) = \varphi(m) \cdot \varphi(n)$. □

Theorem 2.27.

$$\varphi(n) = n \prod_{p\,|\,n} \left(1 - \frac{1}{p}\right),$$

where $\prod_{p\,|\,n} \left(1 - \frac{1}{p}\right)$ *denotes the product over all distinct primes that divide n.*

Proof 1. Assume first that $n = p^k$, where $p \in \mathbb{P}$ and $k \geq 1$. We have $gcd(p^k, u) > 1$, for $1 \leq u \leq p^k$, if and only if $u = \ell p$, with $1 \leq \ell \leq p^{k-1}$. Thus

$$\varphi(p^k) = p^k - p^{k-1} = \left(1 - \frac{1}{p}\right) p^k.$$

In the general case, if n has the prime decomposition $n = p_1^{\alpha_1} \cdot p_2^{\alpha_2} \cdot \ldots \cdot p_l^{\alpha_l}$, then by the previous observation and Theorem 2.26

$$\begin{aligned}
\varphi(n) &= \varphi(p_1^{\alpha_1}) \cdot \varphi(p_2^{\alpha_2}) \cdot \ldots \cdot \varphi(p_l^{\alpha_l}) \\
&= \left(1 - \frac{1}{p_1}\right) p_1^{\alpha_1} \cdot \left(1 - \frac{1}{p_2}\right) p_2^{\alpha_2} \cdot \ldots \cdot \left(1 - \frac{1}{p_l}\right) p_l^{\alpha_l} \\
&= \prod_{p\,|\,n} \left(1 - \frac{1}{p}\right)
\end{aligned}$$

□

Proof 2. Alternative proof to the Theorem 2.26 can be given by the principle of inclusion and exclusion:

$$\begin{aligned}
|A_1 \cup A_2 \cup \cdots \cup A_\ell| = \sum_{j=1}^{\ell} |A_j| - \cdot \sum_{j<k} |A_j \cap A_k| + \cdot \sum_{j<k<m} |A_j \cap A_k \cap A_m| \\
+ \cdots + (-1)^{\ell-1} \cdot |A_1 \cap A_2 \cap \cdots \cap A_\ell|
\end{aligned}$$

where "$|A|$" denotes the number of elements in A. The proof of the principle of principle of inclusion and exclusion is left as an Exercise.

If $n = p_1^{\alpha_1} \cdot p_2^{\alpha_2} \cdot \ldots \cdot p_\ell^{\alpha_\ell}$, $\alpha_j \geq 1$ is again the prime decomposition of n, we apply the principle by setting

$$A_j = \{k \in \{1, 2, \ldots, n\} : p_j \,|\, k\}$$

Then $\varphi(n) = n - |A_1 \cup A_2 \cup \cdots \cup A_\ell|$ and clearly for distinct indices $j_1, \ldots, j_k$

$$\varphi(A_{j_1} \cap A_{j_2} \cap \cdots \cap A_{j_k}) = \frac{n}{p_{j_1} \cdot p_{j_2} \cdot \ldots \cdot p_{j_k}}$$

From this we obtain

$$\begin{aligned}\varphi(n) &= n + (-1)^1 \cdot \sum_{j=1}^{\ell} \frac{n}{p_j} + (-1)^2 \cdot \sum_{1 \leqslant j < k \leqslant \ell} \frac{n}{p_j \cdot p_k} + \cdots + (-1)^\ell \frac{n}{p_1 \cdot p_2 \cdot \ldots \cdot p_\ell} \\ &= n\left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right)\ldots\left(1 - \frac{1}{p_\ell}\right) = n\prod_{p \mid n}\left(1 - \frac{1}{p}\right)\end{aligned}$$

□

Examples. $\varphi(10) = \varphi(2 \cdot 5) = 10\left(1 - \frac{1}{2}\right)\left(1 - \frac{1}{5}\right) = 4.$

$\varphi(60) = \varphi(2^2 \cdot 3 \cdot 5) = 60\left(1 - \frac{1}{2}\right)\left(1 - \frac{1}{3}\right)\left(1 - \frac{1}{5}\right) = 16.$

Theorem 2.28. *We say that $\overline{u} \in \mathbb{Z}_m$ is a unit if it is invertible, i.e., there is $\overline{u}^{-1} \in \mathbb{Z}_m$ with $\overline{u}\,\overline{u}^{-1} = 1$. Furthermore $\overline{u}$ is unit $\Leftrightarrow$ $gcd(u, m) = 1$. Let us define*

$$\mathbb{Z}_m^* := \{\overline{u} \in \mathbb{Z}_m \mid \overline{u} \text{ is a unit}\}.$$

Then $|\mathbb{Z}_m *| = \varphi(m)$.

Proof. $\overline{x} = \overline{u^{-1}} \Leftrightarrow xu \equiv 1 \pmod m \Leftrightarrow xu - my = 1$ for some y. This equation is solvable if and only if $gcd(u, m) = 1$. The last claim follows from the definition of the totient function. □

Definition 2.29. Let $m \geq 2$. We say that the set of integers $U \subset \mathbb{Z}$ is *a reduced residue system* modulo m if U contains exactly one element from each residue class of $\mathbb{Z}_m^*$.

Examples. If $m = 12$, we may choose $U = \{1, 5, 7, 11\}$ or $U = \{13, -5, 29, -1\}$. On the other hand, if $m = 11$, then we may choose e.g. $U = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.

Remark. Thus U is a reduced residue system (mod m), if it induces only those residue classes whose representatives are relatively prime with m.

Theorem 2.30. *Let $\{a_1, a_2, \ldots, a_{\varphi(m)}$ be a reduced residue system modulo m. If $gcd(k, m) = 1$, then also*

$$\{ka_1, ka_2, \ldots, ka_{\varphi(m)}\}$$

is a reduced residue system modulo m.

Proof 1. First of all, $gcd(ka_j, m) = 1$ for any j, since both $gcd(k, m) = 1$ and $gcd(a_j, m) = 1$. Moreover, the numbers are pairwise relatively non-congruent modulo m: if $ka_j \equiv ka_\ell \pmod m$ since $gcd(k, m) = 1$ we may divide the congruence by k and obtain $a_j \equiv$

$a_\ell \pmod{m}$, which means $j = \ell$ by the definition. The claim follows since the size of the set in question is $\varphi(m)$. □

Proof 2. It is enough to show that $f\colon \overline{u} \to \overline{k}\overline{u}$ is a bijection of $\mathbb{Z}_m^*$ onto itself. Now $k \in \mathbb{Z}_m^*$ since $gcd(k, m) = 1$. The map is well-defined since $(ku, m) = 1$ if $= \overline{u} \in Z_m^*$, and then $\overline{k}\overline{u} \in \mathbb{Z}_m^*$. The map is bijection, since obviously it has the inverse map $f^{-1}\colon \overline{u} \to \overline{k}^{-1}\overline{u}$. □

We are now ready to establish Leonhard Euler's generalization of the Fermat's Little Theorem, which is one of the uncountable many results called 'Euler's Theorem'.

Theorem 2.31. *Asume that $m \geq 2$ and $gcd(a, m) = 1$. Then*

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

Proof. By the second proof of Theorem 2.30, the map $f\colon \overline{u} \to \overline{au}$ is a bijection $\mathbb{Z}_m^* \to Z_m^*$. Hence

$$\{\overline{u_1}, \overline{u_2}, \ldots, \overline{u_{\varphi(m)}} = \{\overline{au_1}, \overline{au_2}, \ldots, \overline{au_{\varphi(m)}}\},$$

which implies that $\overline{u_1} \cdot \overline{u_2} \cdot \ldots \cdot \overline{u_{\varphi(m)}} = \overline{a}^{\varphi(m)}(\overline{u_1} \cdot \overline{u_2} \cdot \ldots \cdot \overline{u_{\varphi(m)}})$. Now $\overline{u_1} \cdot \overline{u_2} \cdot \ldots \cdot \overline{u_{\varphi(m)}}$ is invertible, so we may divide it by itself and obtain

$$\overline{a}^{\varphi(m)} = \overline{1},$$

or, equivalently, $a^{\varphi(m)} \equiv 1 \pmod{m}$. □

Example. $p^2 \mid (a^{p^2-p} - 1)$ for all primes p such that if $p \not| a$.

Definition 2.32. If f is a number theoretic function, the sum

$$\sum_{d \mid n} f(d)$$

denotes the sum over all positive divisors of n. Thus

$$\sum_{d \mid n} f(d) \coloneqq \sum_{d \in \{1,2,\ldots,n\}\colon d \mid n} f(d)$$

Example. $\sum_{d \mid 6} f(d) = f(1) + f(2) + f(3) + f(6)$.

Theorem 2.33. *If $n \geq 2$, then $\sum_{d \mid n} \varphi(d) = n$.*

Proof. If $d \geq 1$ and $d \,|\, n$, denote

$$A_d \coloneqq \{k \in \{1, 2, \ldots, n\} \mid gcd(k, n) = d\}$$

By Theorem 2.26, we have $|A_d| = \varphi\left(\frac{n}{d}\right)$.
On the other hand, the sets A_d are disjoint and their union is $\{1, 2, \ldots, n\}$. This

$$n = \sum_{d\,|\,n} |A_d| = \sum_{d\,|\,n} \varphi\left(\frac{n}{d}\right)$$

implies the claim follows as we note that always (Exercise!)

$$\sum_{d\,|\,n} f\left(\frac{n}{d}\right) = \sum_{d\,|\,n} f(d).$$

□

Let's contemplate the previous theorems' implications via a practical example.

2.5 RSA coding system

Public-key cryptography is a special system used for encrypting messages. It is based on the existence of two different types of keys; the receiver's public key can be distributed can be distributed widely without risking security, and any person can send an encrypted message by using it. However, an encrypted message can only be decrypted with the receiver's special private key. Security is guaranteed as long as the receiver keep her/his key secret.

The first really effective and still prototypical coding system of the type described above was the famous RSA coding (R. Rivest, A. Shamir, L. Adleman 1977), variants of which are to this very day still widely used in web browsers, e-mails, VPN servers, chats and other communication channels. Its basic form is as follows:

1. Person A chooses two large primes p and q , (with $p \neq q$) and denotes

   $$\boxed{\text{n = pq}}.$$

2. Next person A picks e such that $1 < e < n$ and $gcd(e, (p-1)(q-1)) = 1$.

3. Finally person A chooses d so that $1 < d < n$ and $de \equiv 1 \pmod{(p\text{ - }1)(q\text{ - }1)}$. This is possible by the condition on e in step 2. Pair (n, d) is the private key of the person A and pair (n, e) is the public key.

4. Now assume that a person B wants to send an encrypted message to A. As a first step B converts the message into a number $m \in \{1, \ldots, n-1\}$. Then B uses the openly available key (n, e) in order to compute the encrypted message $c \in \{1, \ldots, n-1\}$ from the condition

   $$c \equiv m^e \pmod{n}.$$

   .

5. Person B sends the encrypted message c to person A, which can be done openly - in principle, there is no need to hide c !

6. Person A, after receiving c, decrypts the message by simply computing c' from

$$c' \equiv c^d \pmod{n}.$$

Why does the above algorithm work. The answer is simple:

Theorem 2.34. $c' = m \quad (!)$

Proof. We have $\varphi(n) = (p-1)(q-1)$ and, by assumption $de = \ell(p-1)(q-1)+1$ with some integer $\ell \geq 1$.
If $gcd(m,n) = 1$, we obtain by the Euler's Theorem

$$c' \equiv c^d \equiv m^{de} = m^{\varphi(n)\ell} \cdot m \equiv 1 \cdot m \equiv m \pmod{n},$$

hence $c' = m$ as both numbers are among $\{1, \ldots, n-1\}$.
The case $gcd(m,n) > 1$ is left to an exercise. □

Of course, if one want to send a longer message, one may split the message into small enough pieces that can be sent separately by the above method. RSA can be applied into digital signatures as well.

A potential weakness of RSA is related to what is know as the "factoring problem": is it possible to do a prime composition of a given huge number effectively in practice? At the present no such algorithm is known, but on the other hand nobody has proven the non-existence of such algorithms!

Exercise. Why would easy factoring large number make RSA vulnerable?

Exercise. In the situation above, person A may send messages that anybody (with the possession of a public key) may verify 'signed' by A. Think of this in more detail. .

2.6 The Chinese Remainder Theorem

Lemma 2.35.

i) if $gcd(a_i, m) = 1$ for $\forall i \in \{1, 2, \ldots, l\}$, then $gcd(a_1 \cdot a_2 \ldots \cdot a_l, m) = 1$.

ii) If $gcd(a_i, a_j) = 1$, for $i \neq j$, and $a_i \,|\, m$ for $\forall i$, $1 \leq i \leq l$, then $\{a_1 a_2 \ldots a_l\} \,|\, m$.

Proof. Left as an Exercise. □

Now let's delve into the Chinese Remainder Theorem.

Theorem 2.36. *Let $gcd(m_1, m_j) = 1$ for $1 \leq i < j \leq \ell$. The system*

$$\begin{cases} x \equiv b_1 \pmod{m_1} \\ x \equiv b_2 \pmod{m_2} \\ \quad \vdots \\ x \equiv b_\ell \pmod{m_\ell} \end{cases}$$

has always a solution for any values of the b_j:s. All solutions are obtained from the formula

$$x = x_0 + k(m_1 m_2 \dots m_\ell), \quad k \in \mathbb{Z},$$

where x_0 is one of the solutions.

Proof. Assume first that $b_j = 1$, and $b_i = 0$, when $i \neq j$. Denote $M := m_1 \cdot m_2 \cdot \ldots \cdot m_\ell$ and $M_j = \dfrac{M}{m_j}$.

If $x = kM_j$, it fulfils automatically each ith equation in the system for $i \in \{1, 2, \dots, \ell\} \setminus \{j\}$. The jth congruence will hold assuming that $kM_j \equiv 1 \pmod{m_j}$, equivalently if $kM_j - ym_j = 1$ for some y. This linear equation is solvable since $gcd(M_j, m_j) = 1$. Thus we have found the solution in the special case we are considering, call it x_j. In the general case, one checks immediately that

$$x_0 = \sum_{j=1}^{\ell} b_j x_j.$$

solves the system (do it!). If x is another solution, we have $x \equiv x_0 \pmod{m_j}$, for $\forall\ 1 \leq i \leq \ell$. By Lemma 2.35 this implies $M | (x - x_0)$, or equivalently $x = x_0 + k \cdot m_1 \cdot m_2 \cdot \ldots \cdot m_l$. Conversely, all these numbers clearly are solutions. □

Remark. When solving the remainder system lie above, one may actually follow the procedure of the above proof in actual computations. We will see this in the following example.

Problem. Find all numbers $x \geq 1$ such that $\begin{cases} x \equiv 3 \pmod 5 \\ x \equiv 2 \pmod 7 \\ x \equiv 1 \pmod{11} \end{cases}$

Solution. Firstly let us note $gcd(5, 7, 11) = 1$ so the system has a solution.

Now count $M = 5 \cdot 7 \cdot 11 = 385$, and denote $M_1 = \dfrac{385}{5} = 77$, $M_2 = \dfrac{385}{7} = 55$ and $M_3 = \dfrac{385}{11} = 35$. This corresponds to situations where on the right-hand side of the congruence constants are $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$ in that order.

Now let's substitute these values into the congruence equations $kM_j \equiv 1 \pmod{m_j}$ to obtain

$$\begin{cases} x_1 = 77k_1 \equiv 1 \pmod 5 \\ x_2 = 55k_2 \equiv 1 \pmod 7 \\ x_3 = 35k_3 \equiv 1 \pmod{11} \end{cases}$$

In general we may determine the solutions k_j to this kind of Diophantine equations by the Euclidean algorithm. However, in the present situation as the moduli are rather small, we may either try to guess the solutions or note that

$$77k_1 \equiv 1 \pmod 5 \iff 2k_1 \equiv 1 \pmod 5,$$

and we immediately find a solution $k_1 = 3$, which yields $x_1 = 3 \cdot 77 = 231$. In a similar way

$$55k_2 \equiv 1 \pmod 7 \iff -k_2 \equiv 1 \pmod 7,$$

which has e.g. the solution $k_2 = -1$, producing $x_2 = -55$. Finally,

$$35k_3 \equiv 1 \pmod{11} \iff 2k_3 \equiv 1 \pmod{11},$$

which has e.g. the solution $k_3 = 6$, giving $x_3 = 6 \times 35 = 210$.
Next, we obtain a particular solution of the given system as

$$x_0 = 3x_1 + 2x_2 + x_3 = 3 \cdot 231 + 2 \cdot (-55) + 1 \cdot 210 = 793 \equiv 23 \pmod{385},$$

and the general solution for our congruence system is

$$x = 23 + 385k, \quad k \in \mathbb{Z}.$$

Finally, in the original problem one assumed that $x \geq 1$, whence it has exactly the solutions

$$x = 23 + 385k, \quad k \geq 1.$$

Remark. By using the Chinese Remainder Theorem, it is possible to show that if

$$m = m_1 m_2 \dots m_l$$

with $gcd(m_j, m_u) = 1$ for $1 \leqslant j < u \leqslant l$, then

$$\mathbb{Z}_m \cong \mathbb{Z}_1 \otimes \mathbb{Z}_2 \otimes \cdots \otimes \mathbb{Z}_{m_l}$$

where $\cong$ denotes the ring isomorphism and $\otimes$ the direct product of the rings. This is left as an Exercise.

Chapter 3

General polynomial congruences and primitive roots

3.1 General polynomial congruences

The classical "fundamental theorem of algebra" states that a polynomial of degree n at most n roots (in fact, exactly n roots if multiplicities are counted). What is the situation with polynomial congruences of the form

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + ... + a_0 \equiv 0 \pmod{m} \quad ? \tag{3.1}$$

Remark. In number theory we say that roots x_1 and x_2 of the congruence (equation) (3.1) are different if $x_1 \not\equiv x_2 \pmod{m}$.

Example. Let us find roots to the equation $x^2 + 1 \equiv 0 \pmod{5}$. By trial and error we see that the roots are $\{2, 3\}$, and that there no other roots modulo 5.

Example. The equation $x^2 \equiv 1 \pmod{8}$ has 4 different roots: $x \equiv 1, x \equiv 3, x \equiv 5$ and $x \equiv 7 \pmod{8}$.

Remark. The congruence (3.1) is equivalent to the equation

$$\overline{a_n}\,\overline{x}^n + \overline{a_{n-1}}\,\overline{x}^{n-1} + ... + \overline{a_0} = \overline{0}$$

in $\mathbb{Z}_m$. In case $m \in \mathbb{P}$ we no that $\mathbb{Z}_m$ is a field, and may suspect that the analogy with standard polynomial equations is better in this case. This turns out to be the case, as the following basic theorem states.

Theorem 3.1. (Lagrange) *Consider congruence* (3.1) *of degree n, and assume that $m \in \mathbb{P}$ and $m \nmid a_n$. Then the congruence has at most n different* (*non-congruent*) *roots modulo m.*

Remark. The condition $m \nmid a_n$ simply makes sure that the degree of the congruence is exactly n.

Proof. Assume that x_1 is a root of congruence (3.1). By standard division we get

$$p(x) = (x - x_1)q(x) + b$$

where $b = p(x_1) + \in \mathbb{Z}$, and the leading coefficient of q is a_n. Since $m|f(x_1)$ we have $m \mid b$, and our congruence is equivalent to

$$(x - x_1)q(x) \equiv 0 \pmod{m}.$$

If $x_2, x_1 \not\equiv x_2 \pmod{m}$, solves also the equation, we have

$$m \mid (x_2 - x_1)q(x_2).$$

Since $m \in \mathbb{P}$, this implies that $m \mid q(x_2)$, so that all the other roots than x_1 must solve the polynomial congruence

$$q(x) \equiv 0 \pmod{m}.$$

This equation is of degree $n - 1$ and the leading coefficient is not divisible by m. We thus obtain a valid induction step $n - 1 \Rightarrow n$. Furthermore, the claim is true for $n = 0$, since then our congruence takes the form $a_0 \equiv 0 \pmod{m}$, which has no solutions since by assumption $m \not| a_0$. By induction, our proof is complete. □

Corollary 3.2. Let $m \in \mathbb{P}$, and $n < m$. If the congruence

$$a_n x^n + a_{n-1}x^{n-1} + a_0 \equiv 0 \pmod{m}$$

has strictly more than n roots (mod m), then $m \mid a_j$ for all $j = 0, 1, ..., n$, and the congruence is true for all x.

Proof. Assume that the number of roots is greater than n and $m \nmid a_j$ for some $j_0 \in \{0, 1, ..., n\}$. Then we may assume that j_0 is a maximal, and the equation 3.1 is equivalent to the polynomial

$$a_{j_0}x^{j_0} + a_{j_0-1}x^{j_0-1} + ... + a_0 \equiv 0 \pmod{m},$$

where $m \nmid a_{j_0}$. Then by the Lagrange's theorem the number of roots less or equal than j_0, which is a contradiction. Thus $m \mid a_j$ for $\forall j \in \{0, 1, ..., n\}$, and the congruence is identically valid. □

Corollary 3.3. If $p \in \mathbb{P}$ and $d \mid p - 1$, then the congruence

$$x^d \equiv 1 \pmod{p}$$

has $\underline{\text{exactly } d}$ solutions modulo p.

Proof. Write $p-1 = dd'$ and note that

$$x^{p-1} - 1 = (x^d - 1)(x^{(d'-1)d} + x^{(d'-2)d} + ... + x^{(d'-d')d}) = (x^d - 1)g(x) \tag{3.2}$$

where the degree of g is $(d'-1)d = p-1-d$, and the highest order coefficient in g is 1. By Lagrange's theorem, the equation

$$g(x) \equiv 0 \pmod p$$

has at most $p-1-d$ solutions modulo p. On the other hand, FLT implies that $x^{p-1}-1 \equiv 0 \pmod p$ has at least $p-1$ solutions modulo p, namely the numbers $1, 2, ..., p-1$. By identity (3.2) we deduce that $x^d - 1 \equiv 0 \pmod p$ must have at least $p-1-(p-1-d) = d$ different solutions modulo p. On the other hand, Lagrange's theorem verifies that there are no more than d solutions. □

Theorem 3.4. (Wilson) *An integer $p \geq 2$ is a prime if and only if $(p-1)! + 1$ is divisible by p.*

Proof. Cases $p = 1$ and $p = 2$ are easy to check since $1 \,|\, 1$ and $2 \,|\, (1+1)$. Assume then that $p \geq 3$ is a prime. By FLT the polynomial congruence

$$x^{p-1} - 1 - (x-1)(x-2)...(x-p-1) \equiv 0 \pmod p$$

has at least $p-1$ roots: $x = 1, 2, ..., p-1$. On the other hand, its degree is less than $p-1$, whence Corollary (3.2) implies that it holds identically for all $x \in \mathbb{Z}$. Setting $x = 0$ yields

$$-1 = -1 \cdot (-2) \cdot ... \cdot (-(p-1)) \equiv 0 \pmod p,$$

which is equivalent to the claim since p is odd. The converse is left as an exercise. □

If $m \notin \mathbb{P}$, the congruence equation $p(x) \equiv 0 \pmod m$ given with a general modulo m can easily be reduced to solving congruences

$$f(x) \equiv 0 \pmod{p_j^{\alpha_j}},$$

where $m = p_1^{\alpha_1} \cdot p_2^{\alpha_2} \cdot ... \cdot p_l^{\alpha_l}$ is a prime factorization. The proof of this claim is left as an Exercise **??** for the reader (see page **??**).

3.2 Primitive roots

Example. How many different residue classes modulo p the sequence $1, a, a^2, a^3, ...$ may contain? Here $gcd(a,p) = 1$ and p is a prime. Let us choose $a = 2$ and try a few different primes; turns out the maximal number is $p-1$.

$$\begin{array}{ll} \underline{\text{Case } p = 5}: & \underline{\text{Case } p = 7}: \\ 2^0 \equiv 1 \pmod{5} & 2^0 \equiv 1 \pmod{7} \\ 2^1 \equiv 2 \pmod{5} & 2^1 \equiv 2 \pmod{7} \\ 2^2 \equiv 4 \pmod{5} & 2^2 \equiv 4 \pmod{7} \\ 2^3 \equiv 3 \pmod{5} & 2^3 \equiv 1 \pmod{7} \\ 2^4 \equiv 1 \pmod{5} & 2^4 \equiv 2 \pmod{7} \\ 2^5 \equiv 2 \pmod{5} & 2^5 \equiv 1 \pmod{7} \\ 2^6 \equiv 4 \pmod{5} & 2^6 \equiv 4 \pmod{7} \\ \quad\vdots & \quad\vdots \\ \text{4 different} & \text{3 different} \\ \text{residue classes.} & \text{residue classes.} \end{array}$$

Clearly in case of 7 there is a periodic pattern at work. On the other hand, if we replace 2 by 3 as the base of the power, we get again a maximum number of different residue classes, as we can see:

$$\begin{array}{l} \underline{\text{Case } p = 7,} \\ \quad\underline{a = 3}: \\ 3^0 \equiv 1 \pmod{7} \\ 3^1 \equiv 3 \pmod{7} \\ 3^2 \equiv 2 \pmod{7} \\ 3^3 \equiv 6 \pmod{7} \\ 3^4 \equiv 4 \pmod{7} \\ 3^5 \equiv 5 \pmod{7} \\ 3^6 \equiv 1 \pmod{7} \\ \quad\vdots \\ \text{6 different} \\ \text{residue classes.} \end{array}$$

In order to study what happens with general p and different bases a, we shall adopt the following definition.

Definition 3.5. Assume that $m \geq 2$ and $gcd(a, m) = 1$. Then $\text{ord}_m(a)$ is the smallest positive exponent $t \geq 1$ such that $a^t \equiv 1 \pmod{m}$. We then say that ”a belongs to an exponent t modulo m” or ”t is the order of a modulo m”.

Remark. Number $\text{ord}_m(a)$ is well-defined since by Euler's theorem we at least have $a^{\phi(m)} \equiv 1 \pmod{m}$

Example. By the previous example about different residue classes, $\text{ord}_5(2) = 4$, $\text{ord}_7(2) = 3$ and $\text{ord}_7(3) = 6$.

Theorem 3.6. *Assume that $gcd(a, m) = 1$ and $m \geq 2$. Let $t = ord_m(a)$.*

i) For exponents $n \geq 0$ we have: $a^n \equiv 1 \pmod{m} \quad \Leftrightarrow \quad t \,|\, n$.

ii) $t \,|\, \varphi(m)$. *Especially,* $1 \leq t \leq \varphi(m)$.

iii) Let $n_1, n_2 \geq 0$. Then $a^{n_1} \equiv a^{n_2} \pmod{m} \Leftrightarrow t \,|\, (n_1 - n_2)$.

iv) If $n \geq 1$, then $ord_m(a^n) = \dfrac{t}{gcd(n, t)}$.

Proof.

i) Write $n = kt + r, 0 \leq r \leq t - 1$, and assume that $a^n \equiv 1 \pmod{m}$. We get

$$1 \equiv a^n \equiv a^{t(k)} a^r \equiv 1^t a^r \equiv a^r \pmod{m}$$

. By the definition of t, since $0 \leq r \leq t - 1$, we must have $r = 0$.

ii) By Euler's theorem, $a^{\varphi(m)} \equiv 1$ (mod m), so the claim follows from i).

iii) Assume $n_1 \geq n_2 \geq 0$ and $a^{n_1} \equiv a^{n_2}$ (mod m). Since $gcd(a^{n_2}, m) = 1$, we may divide by a^{n_2} to get equivalently $a^{n_1 - n_2} \equiv 1$ (mod m). By i), this takes place exactly when $t \,|\, (n_1 - n_2)$.

iv) Denote $s = \text{ord}_m(a^n)$. By i), s is the smallest integer such that $t \,|\, ns$. This implies that $s = \dfrac{t}{gcd(n, t)}$, which is left as an Exercise.

□

Definition 3.7. An integer a is a *primitive root* modulo m if $gcd(a, m) = 1$ and $\text{ord}_m(a) = \varphi(m)$.

Remark. Assume $p \in \mathbb{P}$ and $gcd(a, p) = 1$. Then:

$$\begin{gathered} a \text{ is a primitive root modulo } m \\ \Leftrightarrow \\ \{1, a, a^2, ..., a^{p-2}\} \text{ is a reduced residue system modulo } p \\ \Leftrightarrow \\ \mathbb{Z}_p^* = \{\overline{a}^k \mid k = 0, 1, ..., p - 2\}. \end{gathered}$$

The statements follows from Definitions (3.7) and (2.29), as the reader may check.

Lemma 3.8. Assume $p \in \mathbb{P}$ and $t \geq 1$. Then

i) If $t \nmid (p - 1)$, $\text{ord}_p(a) \neq t$ for $\forall a$

ii) If $t \,|\, (p-1)$, then there is either 0 or $\varphi(t)$ different $a \pmod{p}$ so that $\text{ord}_p(a) = t$ holds.

Proof. i) As $\varphi(p) = p - 1$, the Theorem 3.1 implies that $\operatorname{ord}_p(a) \,|\, (p-1)$.

ii) If $\operatorname{ord}_p(a) = 1$, we have $a \equiv 1 \pmod{\text{p}}$, so that a belongs to a unique residue class modulo p. We may thus assume that $t \geq 2$ and $t \,|\, (p-1)$. Furthermore, assume that there is $a \in \mathbb{Z}$ so that $gcd(a,p) = 1$ and $\operatorname{ord}_p(a) = t$. Then by the Definition 3.7 and *iii)* of Theorem 3.1, the numbers

$$1, a, a^2, ..., a^{t-1} \tag{3.3}$$

are mutually non-congruent modulo p. Moreover, they all solve the congruence

$$x^t \equiv 1 \pmod{p}, \tag{3.4}$$

since $a^{jt} \equiv (a^t)^j \equiv 1^j = 1 \pmod{p}, j = 0, 1, 2, ...$ Lagrange's theorem now implies that the numbers (3.3) give all solutions of the equation 3.4 modulo p. Hence all x such that $\operatorname{ord}_p(x) = t$ are among the numbers in 3.3. By *iv)* of Theorem 3.6, $\operatorname{ord}_p(a^s) = \dfrac{t}{gcd(t,j)}$. Thus $\operatorname{ord}_p(a^j) = t$ only if $gcd(t,j) = 1$ and among exponents $0, 1, 2, ..., t-1$ there are exactly $\varphi(t)$ of them that satisfy this condition.

□

We are now ready for an important result due to Gauss.

Theorem 3.9. *Assume that $p \in \mathbb{P}$ and let $t \geq 1$ with $t \,|\, (p-1)$. Then*

$$|\{n \in \{1, 2, ..., p-1\} \colon \mathit{ord}_p(n) = t\}| = \varphi(t)$$

Especially, primitive roots modulo p always exist and their number is $\varphi(p-1)$.

Proof. Let $\psi(n) := \#\Big\{n \in \{1, 2, ..., p-1\} \mid \operatorname{ord}_p(n) = t\Big\}$. Then we have

i) $\sum_{t\,|\,p-1} \psi(t) = p - 1$ by *ii)* of Theorem 3.6

ii) $\sum_{t\,|\,p-1} \varphi(t) = p - 1$ by Theorem 3.1

iii) $\psi(t) \leq \varphi(t)$ for all $t \,|\, p-1$ by Lemma 3.8.

What is cool here is that the above is clearly possible only if $\psi(t) = \varphi(t)$ for all divisors t of $p - 1$. □

Example. Let $p = 7$, making $p - 1 = 6$ and $t \,|\, 6 \Leftrightarrow t \in \{1, 2, 3, 6\}$. We compute mod 7:

a	a^2	a^3	a^4	a^5	a^6	$\operatorname{ord}_7(a)$
1	1	1	1	1	1	1
2	4	1	2	4	1	3
3	2	6	4	5	1	6
4	2	1	4	2	1	3
5	4	6	2	3	1	6
6	1	6	1	6	1	2

t	$\varphi(t)$	$\{a \colon \operatorname{ord}_7(a) = t\}$
1	1	$\{1\}$
2	1	$\{6\}$
3	2	$\{2, 4\}$
6	2	$\{3, 5\}$

Corollary 3.10. If $p \in \mathbb{P}$, then $\mathbb{Z}_p^*$ is cyclic; in other words, it is generated by one element only.

Proof. Choose a primitive root a modulo p. Then $\mathbb{Z}_p^* = \{\overline{1}, \overline{a}, \overline{a^2}, ..., \overline{a^{p-2}}\}$. □

One may characterize all $m \geq 2$ for which the $\mathbb{Z}_m^*$ is cyclic. We just state the result without a proof.

Theorem 3.11. *$\mathbb{Z}_m^*$ is cyclic exactly when $m = 2$, $m = 4$, $m = p^e$ or $m = 2p^e$, where $p \in \mathbb{P}$, $e \geq 1$.*

We will next take a look at some applications of primitive roots.

Example. How many roots does the equation $x^{28} \equiv 1 \pmod{71}$ have?

Solution. Choose a primitive root a modulo 71, $1 \leq a \leq 70$. Given $x \in \{1, 2, ..., 70\}$, we may pick a unique $j \in \{0, 1, ..., 69\}$ so that $x \equiv a^j \pmod{71}$. Then

$$x^{28} \equiv 1 \Leftrightarrow a^{28j} \equiv 1 \pmod{71} \Leftrightarrow 70 \,|\, 28j \Leftrightarrow 5 \,|\, 2j,$$

so that $j = 5k$, $k \geq 0$, in which case we must have $0 \leq j \leq 69$, so there are 14 solutions:

$$x \equiv \{1, a^5, a^{10}, ..., a^{65}\} \pmod{71}.$$

□

One may note that to find the exact values of the solutions we would need one primitive root, which is is not necessarily very easy deterministically for a very large p. However, there are quite many of them, and usually $\varphi(p-1)$ is not much smaller than $p-1$. In fact, one may verify that always

$$\frac{n}{\varphi(n)} \leq C \log\log(n)$$

for a real-valued constant C. Thus, when using trial and error method, in practice one expects to need to test only $C' \log_{10}\log_{10}(n)$ values before hitting a primitive root! On the downside, this testing can be rather onerous if we are not able to factorize $p-1$, which might even be impossible for a huge $p-1$. However, again in practise one may apply probabilistic methods to overcome this difficulty.

Example. Primitive roots have a fascinating connection to the lengths of decimal expansions of number $\dfrac{1}{p}$. For example:

$$\begin{array}{ll}
\dfrac{1}{3} = 0,\underline{3}33333... & \Rightarrow \text{Period has a length of 1} \\
\dfrac{1}{7} = 0,\underline{142857}142857... & \Rightarrow \text{Period has a length of 6} \\
\dfrac{1}{11} = 0,\underline{09}0909... & \Rightarrow \text{Period has a length of 2} \\
\dfrac{1}{13} = 0,\underline{076923}076923... & \Rightarrow \text{Period has a length of 6} \\
\dfrac{1}{17} = 0,\underline{0588235294117647}0588... & \Rightarrow \text{Period has a length of 16}
\end{array}$$

Observe that sometimes $\frac{1}{p}$ has a particularly long period in its decimal expansion, like $\frac{1}{7}$ and $\frac{1}{17}$ above. Furthermore, then the length of period is $p-1$! This is explained by the following theorem.

Theorem 3.12. *Let $p \in \mathbb{P}$, $p \geq 7$. The length of the period in the decimal expansion of p is $ord_p(10)$, which reaches its maximal value $p-1$ exactly when 10 is a primitive root modulo p.*

Proof. Let $t = \mathrm{ord}_p(10)$ and denote the length of the decimal expansion of $\frac{1}{p}$ by $j \geq 1$.

1. $j \leq t$: By the definition of t, we have $a := \frac{10^t - 1}{p} \in \mathbb{N}$, so that $1 \leq a \leq 10^t - 1$. Hence
$$\frac{1}{p} = \frac{a}{10^t - 1} = \frac{a}{10^t(1-10^{-t})} = \frac{a}{10^t} + \frac{a}{10^{2t}} + \frac{a}{10^{3t}} + \ldots,$$
which implies that $j \leq t$.

2. $j \geq t$: Write the decimal expansion of $\frac{1}{p}$ in the form
$$\frac{1}{p} = \frac{b}{10^k} + \frac{a}{10^{k+j}} + \frac{a}{10^{k+2j}} + \ldots,$$
where $k \geq 0$ and $b \in \{0, 1, \ldots, 10^k - 1\}$. Thus, b corresponds to a possible initial non-periodic segment and $a \in \{1, 2, \ldots, 10^j - 1\}$ is the period, where $a \neq 0$. By summing up the geometric series we get
$$b10^{-k} + 10^{-k}\frac{a}{10^j - 1} = \frac{1}{p}, \quad \text{or}$$
$$p(b(10^j - 1) + a) = 10^k(10^{j-1} - 1).$$
Since $gcd(p, 10^k) = 1$, it follows that $p \,|\, (10^j - 1)$, i.e. $10^j \equiv 1 \pmod p$. Then $t \,|\, j$ by the *ii*) of the Theorem 3.6, so that $j \geq t$. Hence we conclude $j = t$.

□

One interesting open question presented by Gauss about decimal expansions is that are there infinitely many primes p so that $\frac{1}{p}$ has maximal period in its decimal expansion; or, equivalently, is 10 a primitive root modulo p for infinitely many p? The answer is positive under the so-called generalised Riemann hypothesis (Hooley 1969).

Let us finally note that primitive roots can be used to turn products modulo p into sums modulo $p-1$. In a sense we are here dealing with a kind of a logarithm in $\mathbb{Z}_p^*$, and it often refereed to by "index" or "discrete logarithm". It may be denoted either by

Definition 3.13. Assume that $p \in \mathbb{P}$ and a is a primitive root modulo p. Then
$$\mathrm{dlog}_a(n) \ (= \mathrm{ind}_a(n)) \ = k \quad \text{if} \quad a^k \equiv n \pmod{p.}$$

It it left as an Exercise to demonstrate that if $gcd(n_1, p) = 1 = gcd(n_2, 1)$, then

$$\operatorname{dlog}_a(n_1 n_2) = \operatorname{dlog}_a(n_1) + \operatorname{dlog}_a(n_2) \mod (p-1).$$

Example. Let $a \in \{1, 2, .., 10\}$. Then 2 is a primitive root (mod 11) and the different values of $\operatorname{ind}_2(a)$ modulo 11 are:

a	1	2	3	4	5	6	7	8	9	10
$\operatorname{ind}_2(a)$	0	1	8	2	4	9	7	3	6	5

In principle, discrete logarithm could be even be used to solve some polynomial equations:

Example. Find x such that $3x^3 \equiv 4 \pmod{11}$.

Solution.

$$\begin{aligned} &\operatorname{dlog}_2(3) + 3\operatorname{dlog}_2(x) = \operatorname{dlog}_2(4) \text{ (mod 10)}\\ \Leftrightarrow\ & 8 + 3\operatorname{dlog}_2(x) = 2 \text{ (mod 10)}\\ \Leftrightarrow\ & 3\operatorname{dlog}_2(x) = -6 \equiv 4 \pmod{10}\\ \Leftrightarrow\ & \operatorname{dlog}_2(x) = 8 \text{ (mod 10)}\\ \Leftrightarrow\ & x = 3 \text{ (mod 11)}. \end{aligned}$$

Chapter 4

Quadratic residues and the second degree congruences

We have learned how to solve congruences of 1st degree in the form

$$ax + b \equiv 0 \pmod{m}.$$

Next we would like to tackle the 2nd degree congruences of the form

$$ax^2 + bx + c \equiv 0 \pmod{m}$$

Again the most important case is $m \in \mathbb{P}$. Not so surprisingly, all boils down to finding square roots: just translate the equation into $\mathbb{Z}_p^*$, recall the solution formula 2nd degree equations and that $\mathbb{Z}_p^*$ is a field for all $p \in \mathbb{P}$. Henceforth we arrive into the following fundamental question: for what a is the congruence

$$x^2 \equiv a \pmod{p}$$

is solvable? Equivalently we may ask which elements $\overline{a} \in \mathbb{Z}_p^*$ possess ”a square root”.

Example. In $\mathbb{Z}_7$ we have $\overline{0} = \overline{0}^2$, $\overline{1} = \overline{1}^2$, $\overline{2} = \overline{3}^2$, $\overline{4} = \overline{2}^2$, but $\overline{3}$, $\overline{5}$ and $\overline{6}$ do not have square roots.

Definition 4.1. Let $m \geq 2$ and $gcd(a, m) = 1$. We say that a is a *quadratic residue* modulo m if the equation

$$x^2 \equiv a \pmod{m}$$

has solutions. Otherwise a is a *quadratic nonresidue* (mod m).

We will prove later that the quadratic residue modulo m may be reduced to the fundamental case $m \in \mathbb{P}$. Thus we will in this section consider mostly quadratic residue (mod p), where $p \in \mathbb{P}$.

4.1 Legendre symbol and Euler's criterion

Let us now define 'Legendre symbol', which yields an example of a great definition: it both simplifies computations and the theory, even quiding the theory to the right direction.[1]

Definition 4.2. Let $p \in \mathbb{P}$, $p \geq 3$. Set

$$\left(\frac{a}{p}\right) = \begin{cases} 1, & \text{if } a \text{ is a quadratic residue modulo } p \\ -1, & \text{if } a \text{ is a quadratic nonresidue modulo } p \\ 0, & \text{if } p \mid a. \end{cases}$$

Example. By the previous example in this chapter's beginning, we have

$$\begin{aligned} 1 &= \left(\frac{1}{7}\right) = \left(\frac{2}{7}\right) = \left(\frac{4}{7}\right) \\ -1 &= \left(\frac{3}{7}\right) = \left(\frac{5}{7}\right) = \left(\frac{6}{7}\right) \\ 0 &= \left(\frac{0}{7}\right) \end{aligned}$$

Remark. Note that $\left(\frac{a}{p}\right) = \left(\frac{b}{p}\right)$ if $a \equiv b \pmod p$.

Theorem 4.3. *Let $p \geq 3$, $p \in \mathbb{P}$. The number of quadratic residues and non-residues (mod p) is the same and equals $\frac{p-1}{2}$.*

Proof. $\{0, \pm 1, \pm 2, \ldots, \pm\frac{p-1}{2}\}$ forms a string of consecutive integers, and is hence a complete residue system. This means that quadratic residues are exactly the congruence classes

$$\left\{1^2, 2^2, \ldots, \left(\frac{p-1}{2}\right)^2\right\} \tag{4.1}$$

If $j^2 \equiv k^2 \pmod p$, $j, k \in \left\{1, 2, \ldots, \frac{p-1}{2}\right\}$, then since $j^2 - k^2 = (j+k)(j-k)$ we have either

i) $p \mid j - k$ or

ii) $p \mid j + k$, but this is impossible since $2 \leq j + k \leq p - 1$.

Hence $p \mid j - k$ and $j = k$. We have shown that the numbers in the equation (4.1) noncongruent with each other (mod p), which proves the Theorem 4.3. □

Euler found a very useful way to deduce whether a given integer is a residue of a non-residue.

[1] One needs to be slightly careful with the notation, for the Legendre symbol bores quite a resemblance to the notation of fractions. The Legendre symbol $\left(\frac{b}{p}\right)$ can be read 'b Legendre p'.

Theorem 4.4. (Euler) *Let $p \in \mathbb{P}$, $p \geq 3$. Then, for $\forall a \in \mathbb{Z}$*

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

Proof. If $p \,|\, a$, the claim is clear. Otherwise, by the FLT all $x \in \{1, 2, \ldots, p-1\}$ satisfy

$$x^{p-1} - 1 = (x^{\frac{p-1}{2}} - 1)(x^{\frac{p-1}{2}} + 1) \equiv 0 \pmod{p}$$

Hence any of them satisfy exactly one of the congruences

$$x^{\frac{p-1}{2}} \equiv 1 \pmod{p} \quad \text{or} \tag{4.2}$$

$$x^{\frac{p-1}{2}} \equiv -1 \pmod{p} \tag{4.3}$$

If a is a quadratic residue, $a \equiv k^2 \pmod{p}$ for some k, $p \nmid k$, whence $a^{\frac{p-1}{2}} \equiv k^{p-1} \equiv 1 \pmod{\text{p}}$ by the FLT. In other words, all quadratic residues satisfy (4.2). By Lagrange's theorem, congruence (4.2) can not possess other solutions and hence all quadratic non-residues must satisfy congruence (4.3). □

Corollary 4.5. Assume that $p \geq 3$, $p \in \mathbb{P}$. Then

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

Equivalently

$$\left(\frac{-1}{p}\right) = \begin{cases} 1, & p = 4n+1 \\ -1, & p = 4n-1, n \in \mathbb{N} \end{cases}$$

Proof. By the Euler's criterion, $\left(\frac{-1}{p}\right) \equiv -1^{\frac{p-1}{2}} \pmod{p}$, which clearly implies the claim. □

Example. By the Corollary 4.5 if $p = 4n+1$, there is x such that $p \,|\, (x^2+1)$, while this is not true for primes of the form $p = 4n-1$. Think through this carefully and while the first impression is deceiving, convince yourself of the non-triviality of this result.

Finally, we may establish multiplicativity of Legendre's symbol which is a main reason behind its usefulness.

Theorem 4.6. *Let $p \in \mathbb{P}$, $p \geq 3$. Then for any $a, b \in \mathbb{Z}$, it holds that*

i) $\left(\dfrac{ab}{p}\right) = \left(\dfrac{a}{p}\right)\left(\dfrac{b}{p}\right)$

ii) $\left(\dfrac{a^k}{p}\right) = \left(\dfrac{a}{p}\right)^k$, $k \in \mathbb{Z}$.

Proof. The part *ii)* is a corollary of the part *i)*, so let's prove only the first part. In turn, by Euler's criterion

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

which yields the claim. □

Remark. Let $a = \varepsilon q_1^{\alpha_1} q_2^{\alpha_2} \dots q_l^{\alpha_l}$, where $\varepsilon = \pm 1$ and q_i are distinct primes. Then by the Theorem 4.6

$$\left(\frac{a}{p}\right) = \left(\frac{\varepsilon}{p}\right)\left(\frac{q_1}{p}\right)^{\alpha_1} \dots \left(\frac{q_l}{p}\right)^{\alpha_l}$$

Hence, in order to determine $\left(\dfrac{a}{p}\right)$, one needs to be able to compute

$$\left(\frac{\pm 1}{p}\right) \quad \text{and} \quad \left(\frac{q}{p}\right), \quad p, q \in \mathbb{P}.$$

The first one is taken care by Corollary 4.5, but for the second symbol we will introduce the theory of ”quadratic reciprocity”, which was first used by Euler, Legendre and Gauss in 18th century. Gauss gave the first rigorous proofs of it and published them in 1801 in his famous book *Disquisitiones Arithmeticae* at the age of 24, which revolutionized number theory. Before we take a detailed look of this cornerstone in the history of number theory, we need a new way to check which numbers are residues. This is called 'Gauss's Lemma'.

4.2 Gauss's Lemma

The Gauss's Lemma can be used to determine e.g. $\left(\dfrac{2}{p}\right)$ for general p, and it is a main tool in our proof of quadratic reciprocity theorem. There are two versions of the lemma; the first one takes the form:

Lemma 4.7. Let $p \in \mathbb{P}$, $p \geq 3$ and $gcd(a,p) = 1$. Denote by μ the number of those elements in the sequence

$$\{a, 2a, \ldots, \frac{p-1}{2} \cdot a\}$$

whose absolutely smallest remainders are negative. Then

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

Remark. The "absolutely smallest remainder" of a modulo p is the number r such that $r \equiv a \pmod p$ and $|r|$ is minimal. When $p \geq 3$ is odd, equivalently we choose $r \equiv a \pmod p$ from the set

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

Proof. Let $c_1, c_2, \ldots, c_{\frac{p-1}{2}}$ be the absolutely smallest remainders of $a, 2a, \frac{p-1}{2}$. We obtain

$$1 \cdot 2 \cdot \ldots \cdot \frac{p-1}{2} a^{\frac{p-1}{2}} \equiv (-1)^{\mu} \cdot |c_1| \cdot |c_2| \ldots \cdot |c_{\frac{p-1}{2}}| \pmod p$$

If we check that $\{1, 2, \ldots, \frac{p-1}{2}\} = \{|c_1|, |c_2|, \ldots, |c_{\frac{p-1}{2}}|\}$, we may divide both sides of the congruence by $((p-1)/2)!$ and obtain $a^{\frac{p-1}{2}} \equiv (-1)^{\mu} \pmod p$, whence we obtain the the claim by Euler's criterion.
Since $1 \leq |c_j| \leq \frac{p-1}{2}$, it is enough to show that $|c_j| \neq |c_k|$ if $j \neq k$. Otherwise, we would have $ja \equiv \pm ka \pmod{\text{p}}$, or $p \,|\, (j \pm k)a$ or $p \,|\, (j \pm k)$. Since $|j \pm k| \leq p - 1$, we have $j \pm k = 0$, i.e. $j = k$. □

Remark. Recall that $\lfloor x \rfloor = n$, if $n \leq x \leq n + 1, n \in \mathbb{Z}$, where $\lfloor x \rfloor$ denotes the integer part of $x \in \mathbb{R}$.

Lemma 4.8. $\left(\frac{2}{p}\right) = (-1)^{\frac{p^2-1}{8}}$ for odd primes $p \geq 3$. Equivalently,

$$\left(\frac{2}{p}\right) = \begin{cases} 1, & \text{if } p = 8n \pm 1 \\ -1, & \text{if } p = 8n \pm 3, n \in \mathbb{N} \end{cases}$$

Proof. We choose $a = 2$ in the first version Gauss's Lemma (Lemma 4.7). We have to determine how many negative ones are among the absolutely smallest remainders of the numbers

$$\{2, 4, \ldots, p-1\}$$

The absolutely smallest remainder of $2j$ is positive if $2j \leq \frac{p-1}{2}$, so we have

$$\mu = \frac{p-1}{2} - \left\lfloor \frac{p-1}{4} \right\rfloor$$

Every prime $p \leq 3$ is of the form

$$p = 8k + b, b \in \{1, 2, 5, 7\},$$

so we have

$$\begin{cases} b = 1 & \Rightarrow \mu = 4k - 2k = 2k \equiv 0 \pmod{2} \\ b = 2 & \Rightarrow \mu = 2k + 1 - 2k = 2k + 1 \equiv 1 \pmod{2} \\ b = 5 & \Rightarrow \mu = 4k + 2 - (2k+1) = 2k + 1 \equiv 0 \pmod{2} \\ b = 7 & \Rightarrow \mu = 4k + 3 - (2k+1) = 2k + 2 \equiv 0 \pmod{2} \end{cases}$$

and the claim follows by Gauss' lemma. Especially by checking the different values of b we obtain $\left(\frac{2}{p}\right) = (-1)^{\frac{p^2-1}{8}}$. □

Example. $x^2 \equiv 2 \pmod{113}$ has solutions since $113 = 14 \cdot 8 + 1$. The smallest is $x \equiv 51 \pmod{113}$ found by the trial and error.

Next we introduce the second version of the Gauss's Lemma.

Lemma 4.9. Let $p \in \mathbb{P}$, $p \geq 3$. Assume a is odd and $gcd(a,p) = 1$. Then

$$\left(\frac{a}{p}\right) = (-1)^t, \text{ where } t = \sum_{j=1}^{\frac{p-1}{2}} \left\lfloor \frac{ja}{p} \right\rfloor$$

Proof. When $j \in \{1, 2, \ldots, \frac{p-1}{2}\}$, we have

$$ja = p\left\lfloor \frac{ja}{p} \right\rfloor + r_j, \quad r_j \in \{1, 2, \ldots, p-1\}$$

since $p \nmid a$. Write

$$\{r_1, r_2, \ldots, r_{\frac{p-1}{2}}\} = \{s_1, s_2, \ldots, s_\mu\} \cup \{t_1, t_2, \ldots, t_v\}$$

where $\mu + v = \frac{p-1}{2}$ and $s_u > \frac{p-1}{2}$ and $t_k \leq \frac{p-1}{2}$ for every u, k. Thus the absolutely smallest remainders of the numbers ja, $j \in \{1, 2, \ldots, \frac{p-1}{2}\}$ are the numbers

$$s_1 - p, s_2 - p, \ldots, s_\mu - p, \quad t_1, t_2, \ldots, t_v$$

In the proof of the Lemma 4.7 we checked that

$$\{1, 2, \ldots, \frac{p-1}{2}\} = \{p - s_1, p - s_2, \ldots, p - s_u, t_1, t_2, \ldots, t_v\} \tag{4.4}$$

W next compute (mod 2) using the above fact and the simple observations

$$-1 \equiv 1,\ a \equiv 1 \text{ and } p \equiv 1 \pmod 2:$$

$$\begin{aligned}\sum_{j=1}^{\frac{p-1}{2}} j &\overset{(4.4)}{=} \sum_{u=1}^{\mu}(p-s_u) + \sum_{k=1}^{v} t_k \equiv \mu p + \sum_{u=1}^{\mu} s_u + \sum_{k=1}^{v} t_k \\ &= \mu p + \sum_{j=1}^{\frac{p-1}{2}} r_j = \mu p + \sum_{j=1}^{\frac{p-1}{2}} (ja - p\left\lfloor \frac{ja}{p} \right\rfloor) \\ &\equiv \mu + \sum_{j=1}^{\frac{p-1}{2}} j - \sum_{j=1}^{\frac{p-1}{2}} \left\lfloor \frac{ja}{p} \right\rfloor.\end{aligned}$$

By comparing the left- and right-most elements we see that

$$\mu \equiv \sum_{j=1}^{\frac{p-1}{2}} \left\lfloor \frac{ja}{p} \right\rfloor \pmod 2,$$

and the claim follows from Lemma 4.7. □

4.3 Law of quadratic reciprocity

We are now ready for one of the deepest results of our course. Gauss found the first correct proof of it, and he called the result ”Theorem Aureum”. The result fulfils all the characteristics of a landmark result: it is easy to state, it is surprising and the proof is far from obvious. Moreover, it is useful and opens up new research directions in number theory. The quadratic reciprocity predates Artin’s general reciprocity theorem and Langland’s famous program. Without further ado, let’s state the theorem of quadratic reciprocity.

Theorem 4.10. *Let p, q be different odd primes. Then*

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

In other words, $\left(\frac{p}{q}\right) = \left(\frac{q}{p}\right)$ *except in the case where both p and q are of the form* $4n-1$*; then* $\left(\frac{p}{q}\right) = -\left(\frac{q}{p}\right)$.

Proof. We will apply Lemma 4.9. Denote by S the integer pairs $(x, y) \in \mathbb{Z}^2$ such that

$$1 \le x \le \frac{p-1}{2} \quad \text{and} \quad 1 \le y \le \frac{q-1}{2}$$

so that $|S| = \left(\frac{p-1}{2}\right)\left(\frac{q-1}{2}\right)$. Let further

$$S_1 = \{(x, y) \in S \mid qx > py\}$$
$$S_2 = \{(x, y) \in S \mid qx < py\}$$

Then $S = S_1 \cup S_2$, since we have $qx = py \quad \Rightarrow \quad p \mid x$, whence $(x, y) \notin S$. Let us fix $x \in \{1, 2, \ldots, \frac{p-1}{2}\}$. Then $(x, y) \in S_1$ only if $1 \leq y < \frac{q}{p} \cdot x$, or equivalently, $1 \leq y \leq \frac{q}{p} \cdot x$, as we know now that $yp \neq qx$. Thus the number of such pairs with fixed x is $\left\lfloor \frac{xq}{p} \right\rfloor$, where we also need to observe that that $\left\lfloor \frac{q}{p} \cdot \frac{p-1}{2} \right\rfloor \leq \frac{q-1}{2}$ (Exercise!). Putting all together, we have

$$|S_1| = \sum_{j=1}^{\frac{p-1}{2}} \left\lfloor \frac{jq}{p} \right\rfloor$$

In a similar way

$$|S_2| = \sum_{k=1}^{\frac{q-1}{2}} \left\lfloor \frac{kp}{q} \right\rfloor$$

At this stage Lemma 4.9 yields

$$\begin{aligned} \left(\frac{q}{p}\right)\left(\frac{p}{q}\right) &= (-1)^{|S_1|} \cdot (-1)^{|S_2|} = (-1)^{|S_1|+|S_2|} \\ &= (-1)^{|S|} = (-1)^{\frac{p-1}{2} \cdot \frac{q-1}{2}}. \end{aligned}$$

□

With the aid of the quadratic reciprocity theorem one may quickly compute Legendre symbols.

Example. What is the value of $\left(\frac{2819}{4177}\right)$?

Solution. Both $4177, 2819 \in \mathbb{P}$. Hence by the repeated application of Theorem 4.6, Lemma 4.8, and the quadratic reciprocity (Theorem 4.10), we may compute

$$\begin{aligned} \left(\frac{2819}{4177}\right) &\overset{4.10}{=} \left(\frac{4177}{2819}\right) = \left(\frac{1358}{2819}\right) \\ &\overset{4.6}{=} \left(\frac{2}{2819}\right)\left(\frac{7}{2819}\right)\left(\frac{97}{2819}\right) \\ &\overset{4.8}{=} -1 \cdot \left(\frac{7}{2819}\right)\left(\frac{97}{2819}\right) \\ &\overset{4.10}{=} -1 \cdot -\left(\frac{2819}{7}\right)\left(\frac{2819}{97}\right) \\ &= \left(\frac{5}{7}\right)\left(\frac{6}{97}\right) = \left(\frac{7}{5}\right)\left(\frac{2}{97}\right)\left(\frac{3}{97}\right) \\ &= \left(\frac{2}{5}\right)\left(\frac{2}{97}\right)\left(\frac{97}{3}\right) = \left(\frac{2}{5}\right)\left(\frac{2}{97}\right)\left(\frac{1}{3}\right) \\ &= -1 \cdot 1 \cdot 1 = -1. \end{aligned}$$

Thus 2819 is not a quadratic residue modulo 4177.

Let us then consider the general 2nd degree congruence of the form

$$ax^2 + bx + c \equiv 0 \pmod{p}$$

where $p \geq 3$, $p \in \mathbb{P}$ and $p \nmid a$. Now $\mathbb{Z}_p^*$ is a field and we may compute equivalently

$$\begin{aligned} & \overline{a}\,\overline{x}^2 + \overline{b}\overline{x} + \overline{c} = \overline{0} \qquad\qquad | \cdot 4\overline{a}, \;\; 4\overline{a} \neq \overline{0} \\ \Leftrightarrow & 4\overline{a}^2\overline{x}^2 + 4\overline{a}\overline{b}\overline{x} + 4\overline{a}\,\overline{c} = \overline{0} \\ \Leftrightarrow & (2\overline{ax} + \overline{b})^2 = \overline{b}^2 - 4\overline{a}\,\overline{c} := \overline{\triangle}. \end{aligned} \tag{4.5}$$

Thus a necessary condition for solvability is that the quantity $(\overline{\triangle})$ (the 'discriminant') is a square in $\mathbb{Z}_p^*$, which means that $\overline{\triangle} = \overline{y}^2$ for some $y \in \mathbb{Z}_p^*$. In the positive case we thus obtain

$$(4.5) \Leftrightarrow (2\overline{ax} + \overline{b} - \overline{y})(2\overline{ax} + \overline{b} + \overline{y}) = 0,$$

whence we see that all solutions are

$$\overline{x} = \frac{-\overline{b} \pm \overline{y}}{2\overline{a}}$$

which is a division in the field $\mathbb{Z}_p^*$.
Equality of the roots implies that $\overline{y} = 0$. In other words, we have shown the following theorem.

Theorem 4.11. *Assume that $p \geq 3$, $p \in \mathbb{P}$ and consider the congruence*

$$ax^2 + bx + c \equiv 0 \pmod{p},$$

where we assume that $p \nmid a$. Denoting $\triangle := b^2 - 4ac$ the congruence has

$$\begin{cases} 2 \textit{ solutions, if} & \left(\frac{\triangle}{p}\right) = 1 \\ 1 \textit{ solution, if} & \left(\frac{\triangle}{p}\right) = 0 \\ 0 \textit{ solutions, if} & \left(\frac{\triangle}{p}\right) = -1. \end{cases}$$

Example. Does the equation $x^2 + 12x - 43 \equiv 0 \pmod{151}$ have solutions?

Solution. $151 \in \mathbb{P}$. The discriminant is $\triangle = 12^2 - 4 \cdot 1 \cdot -43 = 316 \equiv 14 \pmod{151}$, and

$$\left(\frac{14}{151}\right) = \left(\frac{2}{151}\right)\left(\frac{7}{151}\right) = 1 \cdot \left(-\left(\frac{151}{7}\right)\right) = -\left(\frac{4}{7}\right) = -\left(\frac{2}{7}\right)^2 = -1.$$

Hence there are no solutions!

We end this chapter by taking a look at the solvability of

$$x^2 \equiv a \pmod{m}$$

where general moduli $m \notin \mathbb{P}$, assuming $gcd(a, m) = 1$. As a side remark, the case, where $gcd(a, m) > 1$, can easily be reduced to this case (Exercise). Recall that we now have a rather complete understanding of the case $m \in \mathbb{P}$, so let's continue with the case $m = p^e$, $p \in \mathbb{P}$, $e \geq 2$.

Lemma 4.12. Let $p \geq 3$, $p \in \mathbb{P}$, $e \geq 2$ and $gcd(a, p) = 1$. The congruence

$$x^2 \equiv a \pmod{p^e}$$

has solutions $\Leftrightarrow \left(\frac{a}{p}\right) = 1$. Then there are two different solutions modulo p^e.

Proof. Necessity: If $p^e \mid (x^2 - a)$, then also $p \mid (x^2 - a)$. Since by assumption $p \not| a$, we must have $\left(\frac{a}{p}\right) = 1$.

Sufficiency: If $\left(\frac{a}{p}\right) = 1$ we know that $x^2 \equiv a \pmod{p}$ has some solution; let it be $x_1^2 \equiv a \pmod{p}$. If x_2 is another solution modulo p, we have either $x_1^2 \equiv x_2^2 \pmod{p}$ or $p \mid (x_1^2 - x_2^2)$, and hence $x_2 \equiv \pm x_1 \pmod{p}$. Therefore in the case $e = 1$ there are exactly two solutions.

We proceed now by induction: assume that $e \geq 1$ and

$$x^2 \equiv a \pmod{p^e} \tag{4.6}$$

has exactly two solutions, x_1 and x_2. Now $-x_1$ is a solution too and $-x_1 \not\equiv x_1 \pmod{p^e}$, whence we must have $x_2 \equiv -x_1 \pmod{p^e}$. If x now solves the congruence

$$x^2 \equiv a \pmod{p^{e+1}}, \tag{4.7}$$

then x also solves the (4.6), i.e. we must have $x \equiv \pm x_1 \pmod{p^e}$. In other words, the only possible solutions (mod p^{e+1}) to congruence (4.7) are are found in the set

$$x = \pm x_1 + tp^e, \quad t \in \{0, 1, 2, \ldots, p-1\}$$

Now

$$x^2 - a \equiv (\pm x_1 + tp^e)^2 - a \equiv {x_1}^2 \pm 2tx_1p^e - a \pmod{p^{e+1}},$$

or equivalently,

$$2tx_1 + (x_1^2 - a)/p^e \equiv 0 \pmod{p}.$$

For both signs this congruence has unique solution $t \in \{0, 1, 2, \ldots, p-1\}$ since $gcd(2x_1, p) = 1$. □

We also need to consider the case $m = 2^e$ which is slightly more technical (this is kind of extra material for the lectures).

Lemma 4.13. Let $gcd(a, 2) = 1$ and $e \geq 1$. The congruence

$$x^2 \equiv a \pmod{2^e} \tag{4.8}$$

has solutions if and only if

i) $e = 1$, whence there is one solution.

ii) $e = 2$, and $a \equiv 1 \pmod 4$ whence there are two solutions.

iii) $e \geq 3$ and $a \equiv 1 \pmod 8$, when the are four solutions.

Proof. We give a rather condensed account here – no need to worry about the missing details, they are easy.

i) Clearly $x \equiv 1 \pmod 2$ is the only solution.

ii) x must be odd $\Rightarrow x \equiv \pm 1 \pmod 4 \Rightarrow a \equiv 1 \pmod 4$. Clearly all the solutions are $x \equiv \pm 1 \pmod 4$.

iii) The statement follows if we show by induction the following claim for $e \geq 3$:

Claim: If $e \geq 3$ and $a \equiv 1 \pmod 8$, the congruence (4.6) has four solutions modulo 2^e, of which exactly 2 solve[2] the same congruence with respect to the moduli 2^{e+1}.

Before proving the claim, note first that $a \equiv 1 \pmod 8$ is a necessary requirement: now $a \equiv x^2 \pmod{2^e} \quad \Rightarrow \quad a \equiv x^2 \pmod 8$ and as x is odd, we have

$$x^2 = (2n + 1)^2 = 4n(n + 1) + 1$$

and here either $2 \,|\, n$ or $2 \,|\, (n + 1)$.

Proof of the claim:

Base case, e= 3: $x^2 \equiv 1 \pmod 8$ has four solutions: $x = \pm 1$ or $x = \pm 3$, which we can figure out by trial. Given $a \equiv 1 \pmod 8$ we have either $a \equiv 1$ or $a \equiv 9 \pmod{16}$. The roots $x = \pm 1$ solve t $x^2 \equiv a \pmod{16}$ in the former case and the roots $x = \pm 3$ respectively in the latter case. Thus the claim holds for $e = 3$.

Induction step: Assume that claim holds for some $e \geq 3$. In other words, we assume that of numbers $\{0, 1, 2, \ldots, 2^e\}$ only x_1, x_2, x_3 and x_4 solve (4.8), and that (say) x_1 and x_2 solve $x^2 \equiv a \pmod{2^{e+1}}$. All the different solutions of this latter congruence are found among the numbers $\{x_1, x_2, x_3, x_4, x_1 + 2^e, x_2 + 2^e, x_3 + 2^e, x_4 + 2^e\}$ (why?). By our assumption x_3 and x_4 do not solve the congruence. Then also $x_3 + 2^e$ and $x_4 + 2^e$ do not because e.g.

$$(x_3 + 2^e)^2 \equiv {x_3}^2 + 2^{e+1} + 2^{2e} \equiv {x_3}^2 \pmod{2^{e+1}} \tag{4.9}$$

[2] note that this is independent of the representative: the numbers $x + k2^e$ solve the congruence (4.8) (mod 2^{e+1}) simultaneously for all $k \in Z$ by similar computation to (4.9) !

Thus the equation has exactly the solutions $x_1, x_1 + 2^e, x_2$ and $x_2 + 2^e$. The rest of the claim follows if we check that exactly one of x_1 and $x_1 + 2^e$ solve the equation

$$x^2 \equiv a \pmod{2^{e+2}} \tag{4.10}$$

since then the same claim is true for x_2 and x_2+2^e. This follows (why?) by observing that

$$(x_1 + 2^e)^2 - (x_1)^2 \equiv 2^{2e} + 2^{e+1}x_1 \equiv 2^{e+1} \pmod{2^{e+2}}.$$

□

Lemma 4.14. Let $m = m_1, m_2, \ldots m_\ell$, where m_i are pairwise relative primes and $gcd(a, m) = 1$. Then the number of solutions to the congruence equation

$$x^2 \equiv a \pmod{m} \tag{4.11}$$

equals $N_1 \cdot l_2 \cdot \ldots \cdot N_\ell$, where N_j is the number of solutions to the equation

$$x^2 \equiv a \pmod{m_j} \tag{4.12}$$

Proof. Let $x_{k,j}$ be the different of solutions of congruence (4.12) modulo m_j, $1 \le k \le N_j$. Then x solves (4.11) if and only if

$$\begin{cases} x \equiv x_{k_1,1} \pmod{m_1} \\ x \equiv x_{k_2,2} \pmod{m_2} \\ \quad\vdots \\ x \equiv x_{k_\ell,\ell} \pmod{m_\ell} \end{cases}$$

for some choice of the indices $k_1 \in \{1, 2, \ldots, N_1\}$, $k_2 \in \{1, 2, \ldots, N_2\}, \ldots, k_\ell \in \{1, 2, \ldots, N_j\}$. By the Chinese Remainder Theorem (!) for each such choice there is exactly one solution x modulo m that solves this system of linear equations. Thus the number of different solution mod m is $N_1 \cdot N_2 \cdot \ldots \cdot N_\ell$. □

We may now apply all Lemmas 4.12, 4.13 and 4.14 together and deduce the following general result:

Theorem 4.15. *Let $gcd(a, m) = 1$, $m \geq 2$.*

i) The congruence

$$x^2 \equiv a \pmod{m} \tag{4.13}$$

has solutions if and only if $\left(\dfrac{a}{p}\right) = 1$ *for all odd primes p that divide m and, moreover, $a \equiv 1 \pmod 4$ if $4 \mid m$ and $a \equiv 1 \pmod 8$ if $8 \mid m$.*

ii) Let $m = 2^e \cdot p_1^{e_1} \cdot p_2^{e_2} \cdot \ldots \cdot p_l^{e_l}$ where p_i are different odd primes. Assume that the congruence (4.13) has solutions. Then their number is 2^{l+E}, where

$$E = \begin{cases} 0, & \textit{if } e \leq 1 \\ 1, & \textit{if } e = 2 \\ 2, & \textit{if } e \geq 3. \end{cases}$$

Example. How many solutions does the congruence $x^3 \equiv 113 \pmod{196}$ have?

Solution. Now $196 = 2^2 \cdot 7^2$.

- Congruence $x^2 \equiv 113 \pmod{7^2}$ has solutions since $\left(\dfrac{113}{7}\right) = \left(\dfrac{1}{7}\right) = 1$.
- Congruence $x^2 \equiv 113 \pmod 4$ has solutions since $113 \equiv 1 \pmod 4$.
- By Theorem 4.15, the total number of solutions is $2^{1+1} = 4$.

Chapter 5

Diophantine approximation — from algebraic numbers to trisection of angles by a ruler and a compass

Example. If $x \in]0,1[$ and $q \in \mathbb{N}$, we may consider the division

$$[0,1[= \bigcup_{k=1}^{q} [\tfrac{k-1}{q}, \tfrac{k}{q}[$$

and deduce that one may approximate x by a rational number with denominator $p \leq q$, where the error is $\varepsilon \leq \frac{1}{q}$.

Is this the best one can do, however? How does the result reflect the nature of x? For example, if x is algebraic (meaning it satisfies a polynomial equation with integral coefficients), does it show up in the results?

We start with a fundamental observation discovered by a mathematician Peter Dirichlet, the "Dirichlet's Lemma". The proof of the lemma introduced the "Dirichlet's (drawer) box principle" or "Dirichlet's pigeonhole principle" for the first time.

Lemma 5.1. If $x \in \mathbb{R}$, $t \in \mathbb{N}$, there are integers n, m such that $1 \leq m \leq t$ and

$$|mx - n| \leq \frac{1}{t+1}$$

Proof. We may assume that $x > 0$. Write

$$I_k = [\frac{k-1}{t+1}, \frac{k}{t+1}],$$

so that $[0,1[= \bigcup_{k=1}^{t+1} I_k$.

Now consider the decimal parts

$$x - \lfloor x \rfloor, 2x - \lfloor 2x \rfloor, ..., tx - \lfloor tx \rfloor.$$

1. If some of the decimal parts is in I_1, say $kx - \lfloor kx \rfloor \in I_1$, we may choose $m = k$, $n = \lfloor kx \rfloor$.

2. If one of the decimal parts lies in I_{t+1}, say $kx - \lfloor kx \rfloor \in I_{t+1}$, then

$$1 - \frac{1}{t+1} \leq kx - \lfloor kx \rfloor < 1 \Rightarrow -\frac{1}{t+1} \leq kx - (\lfloor kx \rfloor + 1) < 0,$$

whence we may choose $m = k$, $n = \lfloor kx \rfloor + 1$.

3. If neither (1) nor (2) holds, then by the pigeonhole principle two of the decimal parts lie in the same interval I_j. Especially, we may pick $1 \leq k_1 < k_2 \leq t$ so that

$$\begin{aligned} &|k_2x - \lfloor k_2x \rfloor - (k_1x - \lfloor k_1x \rfloor)| < \frac{1}{t+1} \\ \Leftrightarrow\, &|(k_2 - k_1)x - (\lfloor k_2x \rfloor - \lfloor k_1x \rfloor)| < \frac{1}{t+1} \end{aligned}$$

Our final choice is then $m = k_2 - k_1$ and $n = \lfloor k_2x \rfloor - \lfloor k_1x \rfloor$.

□

Corollary 5.2. i) Let $x \in \mathbb{R}$. For every $t \in \mathbb{N}$ there are $p, q \in \mathbb{Z}$ such that $1 \leq q \leq t$ and

$$\left| x - \frac{p}{q} \right| \leq \frac{1}{(t+1)q} < \frac{1}{q^2}$$

ii) Let $x \in \mathbb{R} \backslash \mathbb{Q}$. There are infinitely many different rationals $\frac{p}{q}$ with $\left| x - \frac{p}{q} \right| \leq \frac{1}{q^2}$.

Proof. i) By Lemma 5.1, for any positive integer $t \geq 2$ there are integers $p_t, q_t \in \mathbb{Z}$, $1 \leq q_t \leq t$ with $|q_tx - p_t| \leq \frac{1}{t+1} \Rightarrow \left| x - \frac{p_t}{q_t} \right| \leq \frac{1}{(t+1)q_t} < \frac{1}{q_t^2}$.

ii) Apply the part *i)* to each t to obtain

$$\left| x - \frac{p_t}{q_t} \right| \leq \frac{1}{(t+1)q_t} < \frac{1}{q_t^2}.$$

Since the approximation error $\varepsilon \leq \frac{1}{t+1} \to 0$ as $t \to \infty$, there must be infinitely many different values for $\frac{p_t}{q_t}$ because $x \notin \mathbb{Q}$!

□

Remark. The estimate in the Corollary 5.2 cannot in general be improved with regards the power in q: we claim that

$$\left| \sqrt{2} - \frac{p}{q} \right| > \frac{1}{4q^2}$$

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

Namely, assume that $\left|\sqrt{2}-\frac{p}{q}\right| \leq \frac{1}{4q^2}$, with $q, p \in \mathbb{Z}$ and $q \geq 1$. Then $p \geq 1$ (why?) and we have

$$\sqrt{2}+\frac{p}{q} \leq 2\sqrt{2}+\left|\frac{p}{q}-\sqrt{2}\right| \leq 2\sqrt{2}+\frac{1}{4} < 4$$

Thus, it follows that

$$\left|2-\frac{p^2}{q^2}\right| = \left|\sqrt{2}-\frac{p}{q}\right| \cdot \left|\sqrt{2}+\frac{p}{q}\right| \leq \frac{1}{4q^2}\Big(\sqrt{2}+\frac{p}{q}\Big) < \frac{1}{q^2}.$$

This implies that $|2q^2-p^2| < 1$, so that $2q^2 = p^2$, which is a contradiction. The proof is complete.

5.1 Pell's equation

Pell's equation is any Diophantine equation of the form

$$x^2 - Dy^2 = 1, \qquad D \in \mathbb{N}, \sqrt{D} \notin \mathbb{Q}$$

Pell's equation has a long history. It was studied by Babylonians, Archimedes, Chinese, Indian and Arab mathematicians, as well as Fermat, Euler, Lagrange, Wallis, Brouncker, along with many others. Even Archimedes, when posing his famous "Sun God's Cattle" problem, apparently knew that it leads to a Pell's equation with huge a minimal solution. Euler named the equation after the English mathematician John Pell, while Pell's only contribution to the subject was the partial publication of Wallis and Brouncker's results. The firstcomplete solution was given by Joseph Lagrange, and a simpler approach that we present is due to Dirichlet.

Let's take a closer look of the equation itself. By trial and error we find that the smallest positive integer solutions are

$$\begin{aligned}
&x^2 - 2y^2 = 1 : x_1 = 3\\
&x^2 - 3y2 = 1 : x_1 = 2\\
&x^2 - 5y^2 = 1 : x_1 = 9\\
&\vdots \qquad\qquad\quad \vdots\\
&x^2 - 61y^2 = 1 : x_1 = 1766319049!
\end{aligned}$$

Thus it is not obvious that Pell's equation is always solvable in positive integers, but we will soon establish a positive result. Before that we will need an auxiliary result.

Lemma 5.3. Let $D \geq 2$, $\sqrt{D} \notin \mathbb{Q}$. Then there is $k \in \mathbb{Z} \setminus \{0\}$ such that $|k| \leq 2\sqrt{D}+1$ and the equation

$$x^2 - Dy^2 = k$$

has infinitely many solution pairs $(x, y) \in \mathbb{N}^2$.

Proof. Choose by the Dirichlet's lemma for every $t \geq 1$ positive integers x_t, y_t with $1 \leq y_t \leq t$ and

$$\left| y_t\sqrt{D} - x_t \right| \leq \frac{1}{t+1}.$$

Then $x_t \leq y_t\sqrt{D} + 1 \leq t\sqrt{D} + 1$ so that

$$\begin{aligned} \left| x_t^2 - y_t^2 D \right| &= \left| x_t + \sqrt{D}y_t \right| \left| x_t - \sqrt{D}y_t \right| \\ &\leq \left| t\sqrt{D} + 1 + t\sqrt{D} \right| \cdot \frac{1}{t+1} \\ &\leq \frac{t(1+2\sqrt{D})}{t} = 2\sqrt{D} + 1 \end{aligned}$$

Thus the number $|x_t^2 - y_t^2 D|$ takes only finitely many values, whence at least one of them, call it k, occurs infinitely many times (by a variant of the pigeonhole principle). Moreover, the corresponding pairs (x_t, y_t) form an infinite set since the quantity $y_t\sqrt{D} - x_t$ tends to 0 as $t \to \infty$. □

The following observation is useful for our further purposes.

Lemma 5.4. Let $D \geq 2$, $\sqrt{D} \in \mathbb{Q}$. Then, if

$$a_1 + b_1\sqrt{D} = a_2 + b_2\sqrt{D}$$

for integers a_1, a_2, b_1 and b_2, it follows that $a_1 = a_2$ and $b_1 = b_2$.

Proof. If e.g. $b_1 = b_2$, we obtain $\sqrt{D} = \dfrac{a_1 - a_2}{b_2 - b_1} \in \mathbb{Q}$, which is contradiction. □

It is useful to note that if $x^2 - Dy^2 = 1$, then also $|x|^2 - D\,|y|^2 = 1$, so it is enough to determine the positive solutions of Pell's equation. Namely, then all solutions are given by

$$\{(\pm x, \pm y) \mid (x, y) \text{ is a positive solution to } x^2 - Dy^2 = 1.\}$$

Theorem 5.5. *Assume that $D \in \mathbb{N}$ is such that $\sqrt{D} \notin \mathbb{Q}$. Then*

i) Pell's equation $x^2 - Dy^2 = 1$ has always positive solutions in integers.

ii) If (x_1, y_1) is the smallest positive solution to the Pell's equation, (i.e., $y_1 \geq 1$ is minimal), then all positive solutions of the equation are obtained from the formula

$$x_n + y_n\sqrt{D} = (x_1 + y_1\sqrt{D})^n, \quad n \geq 1.$$

Consequently, Pell's equation has always infinitely many solutions.

Proof. We use Lemma 5.4 to pick $k \neq 0$ so that the equation $x^2 - Dy^2 = k$ has infinitely many solutions. This enables us to choose (just consider $k^2 + 1$ different positive solutions

– again by the pigeonhole principle!) integers $x_1, x_2, y_1, y_2 \geq 1$ so that $y_1 \neq y_2$ and

$$\begin{cases} x_1^2 - Dy_1^2 = k = x_2^2 - Dy_2^2 \\ x_1 \equiv x_2 \pmod{|k|} \\ y_1 \equiv y_2 \pmod{|k|} \end{cases}$$

Let us denote $z = x_1x_2 - Dy_1y_2$ and $w = x_1y_2 - x_2y_1$ so that

$$(x_1 - \sqrt{D}y_1)(x_2 + \sqrt{D}y_2) = z + w\sqrt{D}$$

We have

$$\begin{aligned} (x_1 + y_1\sqrt{D})(x_1 - y_1\sqrt{D}) &= x_1^2 - Dy_1^2 = k \quad \text{and} \\ (x_2 + y_2\sqrt{D})(x_2 - y_2\sqrt{D}) &= x_2^2 - Dy_2^2 = k. \end{aligned}$$

By multiplying these equalities side by side it follows that

$$\begin{aligned} & k^2 = [(x_1 - y_1\sqrt{D})(x_2 + y_2\sqrt{D})] \cdot [(x_1 + y_1\sqrt{D})(x_2 - y_2\sqrt{D})] \\ \Leftrightarrow\ & k^2 = (z + w\sqrt{D})(z - w\sqrt{D}) \\ \Leftrightarrow\ & k^2 = z^2 - Dw^2 \end{aligned} \tag{5.1}$$

Now $w = x_1y_2 - x_2y_1 \equiv x_1y_1 - x_1y_1 = 0 \pmod{|\mathrm{k}|}$ so that $k \mid w$. By (5.1) we have $k^2 \mid z^2$, whence $k \mid z$. We define integers x, y by setting

$$x := \frac{z}{k}, \quad y := \frac{w}{k}$$

and observe $x^2 - Dy^2 = 1$. We have thus found a nontrivial solution of Pell's equation as soon as we check that $y \neq 0$, or equivalently that $w \neq 0$. However, if $w = 0$, then $\frac{x_1}{y_1} = \frac{x_2}{y_2}$ so that

$$\frac{1}{y_1^2} = \frac{x_1^2}{y_1^2} - k = \frac{x_2^2}{y_2^2} - k = \frac{1}{y_2^2}$$

whence $y_1 = y_2$ since both are greater or equal to 1. This is a contradiction, so $w \neq 0$.

Now that we know that there are non-trivial solutions, let (x_1, y_1) be the smallest positive solution, thus $x_1^2 - Dy_1^2 = 1$, $x_1 \geq 1$, $y_1 \geq 1$, and y_1 is minimal. We define the numbers (x_n, y_n) by setting

$$x_n + y_n\sqrt{D} = (x_1 + y_1\sqrt{D})^n, \quad n \geq 1$$

Then clearly $(x_n - y_n\sqrt{D}) = (x_1 - y_1\sqrt{D})^n$ (note the use of Lemma 5.4 in the definition of x_n and y_n) and we may compute

$$\begin{aligned} x_n^2 - Dy_n^2 &= (x_n + y_n\sqrt{D})(x_n - y_n\sqrt{D}) \\ &= (x_1 + y_1\sqrt{D})^n(x_1 - y_1\sqrt{D})^n \\ &= (x_1^2 - y_1^2\sqrt{D})^n = 1^n = 1. \end{aligned}$$

Thus (x_n, y_n) solves Pell's equation and from the definition it follows that

$$1 \leq y_1 < y_2 < y_3 < \ldots$$

It remains to check that the sequence (x_n, y_n), $n \geq 1$ gives all the positive solutions. Assume to the contrary that this is not true. Then there are $x, y \in \mathbb{Z}$ with $x, y \geq 1$ and $y > y_1$ so that

$$\begin{aligned}
&x^2 - Dy^2 = 1,\\
&x + y\sqrt{D} \neq x_n + y_n\sqrt{D}, \quad \forall n \geq 1
\end{aligned}$$

Now $x = \sqrt{1 + Dy^2} > \sqrt{1 + Dy_1^2} = x_1$. Thus

$$x + y\sqrt{D} > x_1 + y_1\sqrt{D}$$

On the other hand, $x_1 \geq 2$ and $x_n \geq x_1^n$. Thus $x_n + y_n\sqrt{D} \to \infty$ as $n \to \infty$, so we may choose $n \geq 1$ so that

$$(x_1 + y_1\sqrt{D})^n < x + y\sqrt{D} < (x_1 + y_1\sqrt{D})^{n+1}$$

If we multiply the above inequality by $(x_1 - y_1\sqrt{D})^n = (x_1 + y_1\sqrt{D})^{-n}$, we obtain

$$1 < (x + y\sqrt{D})(x_1 - y_1\sqrt{D})^n < x_1 + y_1\sqrt{D}. \tag{5.2}$$

Let us denote $(x + y\sqrt{D})(x_1 - y_1\sqrt{D})^n =: a + b\sqrt{D}$. Then

$$\begin{aligned}
a^2 - Db^2 &= (x + y\sqrt{D})(x_1 - y_1\sqrt{D})^n \cdot (x - y\sqrt{D})(x_1 + y_1\sqrt{D})^n && (5.3)\\
&= (x^2 - Dy^2)(x_1^2 - Dy_1^2)^n = 1. && (5.4)
\end{aligned}$$

Thus (a, b) is a solution as well. If we show that $a, b \geq 1$, then (5.3) implies that (a, b) gives a smaller solution than (x_1, y_1), which is the desired contradiction.

By (5.3) $a + b\sqrt{D} > 1$, whence $a - b\sqrt{D} = \dfrac{1}{a + b\sqrt{D}} \in]0, 1[$. Thus

$$\begin{aligned}
a &= \frac{1}{2}((a + b\sqrt{D}) + (a - b\sqrt{D})) > 0\\
b &= \frac{1}{2\sqrt{D}}((a + b\sqrt{D}) - (a - b\sqrt{D})) > 0
\end{aligned}$$

so that $a, b \geq 1$. The proof is complete. □

5.2 Algebraic numbers

Definition 5.6. A real number x is *algebraic* if it satisfies an equation of the form

$$a_n x^n + a_{n-1}x^{n-1} + ... + a_0 = 0,$$

where $a_0, a_1, ..., a_n \in \mathbb{Z}$ and $a_n \neq 0$. The smallest possible n is the algebraic degree of n. If $x \in \mathbb{R}$ is not algebraic, it is *transcendental.*

The next lemma follows directly from the definitions.

Lemma 5.7. i) $\mathbb{Q} = \{x \in \mathbb{R} \mid x \text{ is an algebraic number of degree 1 }\}$

ii) Algebraic numbers of degree $n \geq 2$ are irrational.

The proof of the following lemma will be an Exercise.

Lemma 5.8. i) A $x \in \mathbb{R}$ is a algebraic of degree 2 if and only if it can be written in the form

$$x = a + b\sqrt{D},$$

where $D \in \mathbb{N}$, $\sqrt{D} \notin \mathbb{Q}$, $a, b \in \mathbb{Q}$, $b \neq 0$.

ii) The set of algebraic numbers is countable.

The next theorem is not very difficult to prove, but we skip it and only sketch the proof at lectures.

Theorem 5.9. *Denote by $\overline{\mathbb{Q}}$ all the algebraic reals. Then $\overline{\mathbb{Q}} \subset \mathbb{R}$ is a subfield of $\overline{\mathbb{Q}}$. In particular, it is invariant under sums and products.*

Let us illustrate the Theorem 5.9 with a couple examples.

Example. $\sqrt{2}$ is algebraic since $\sqrt{2}$ satisfies the equation $x^2 - 2 = 0$. Also $\sqrt[3]{2}$ is algebraic, and then by the Theorem 5.9 $y = \sqrt{2} + \sqrt[3]{2}$ is algebraic as well. To prove this directly, note that $(y - \sqrt{2})^3 = 2$ so that $y^3 - 3y^2\sqrt{2} + 3y(\sqrt{2})^2 - (\sqrt{2})^3 = 2$, or $y^3 + 6y - 2 = \sqrt{2}(3y^2 + 2)$. Hence $(y^3 + 6y - 2)^2 = 2(3y^2 + 2)^2$, which leads to an equation with integer coefficients for y.

Example. If x is obtained from rationals by finitely many operations of multiplications, additions, divisions and taking roots, then x is algebraic. This follows by iterating the Theorem 5.9 and noting that if $x > 0$ is algebraic, say

$$a_n x^n + a_{n-1}^{n-1} + ... + a_0 = 0,$$

where $a_j \in \mathbb{Z}$ and $a_n \neq 0$, then for any integer $k \geq 0$, $x^{\frac{1}{k}}$ is also algebraic, since it satisfies

$$a_n y^{kn} + a_{n-1} y^{k(n-1)} + ... + a_0 = 0.$$

Example. $\dfrac{\sqrt[15]{\sqrt[16]{17} + 1} - \sqrt[19]{20}}{\sqrt[13]{14} + \sqrt[14]{13}}$ is algebraic!

One may solve 3rd degree equations in general by rational and root operations (see Theorem 5.10 below), and the same holds true for 4:th degree equations. There is an interesting history behind:
Remarks.

- Already Babylonians in the 3000 BC knew how to solve 2nd degree equations.

- The first person to find a solution for 3rd degree equations was Scipione del Ferro, the professor of Bologna University, in 1515. He revealed the solution to his student Antonio Fior, while around 1535 another Italian mathematician Niccolò Tartaglia found independently an almost general solution. A public competition about the honor of finding the result ensued, where Tartaglia triumphed over Fior. Later a renounced Italian maathematician Gerolamo Cardano obtained the solution from Tartaglia in a form of a poem, and published it in his book "Ars Magna", which spiked a heated and long dispute between Tartaglia and Cardano, for Tartaglia claimed Cardano having sworn not to reveal it! Today, however, the coveted formula is referred as "Cardano-Tartaglia formula" to honor them both.

- The solution to the general quartic (4th degree) equation was found by Ludovico Ferrari, a student of Cardano.

- Solution of the 3rd degree equations almost forces one to use complex numbers! Cardano even made formal computations of the form $(5+\sqrt{-15})(5-\sqrt{-15}) = 40$.

- The mathematician Niels Abel (after whom the Abel Prize of Mathematics is named) showed in 1824 that a general quintic (5th degree) equation cannot be solved by radicals (root operations). Earlier proof by the Italian mathematician Paolo Ruffini was not complete, while ingenious. The Galois theory, created by the French mathematician Evariste Galois in 1830 gave an explanation for all these phenomena.

- Dividing by a_3 the general 3rd degree equation $a_3x^3+a_2x^2+a_1x+a_0 = 0$ it reduces to the form $x^3 + a_2x^2 + a_1x + a_0 = 0$. By substituting here $x = y - a_2/3$, this reduces to the form where also $a_2 = 0$. Thus, in order to find a general solution it is enough to consider 3rd degree equations of the form $x^3 + px + q = 0$.

Theorem 5.10. *The 3rd degree equation of the form*

$$x^3 + px + q = 0$$

has the roots

$$u_0 + v_0, \rho u_0 + \rho^2 v_0 \text{ and } \rho^2 u_0 + \rho v_0 \tag{5.5}$$

where $\rho = e^{\frac{2\pi i}{3}}$ *and*

$$u_0 = \sqrt[3]{-\frac{q}{2} + \sqrt{-D}}, \qquad v_0 = \sqrt[3]{-\frac{q}{2} - \sqrt{-D}},$$

and the cubic roots are chosen so that $u_0v_0 = -\frac{p}{3}$. *Above* D *the* discriminant: $D := -\Big((q/2)^2 + (p/3)^3\Big)$.

Proof. The idea is to substitute $x = u + v$, whence the equation takes the form

$$u^3 + v^3 + 3uv(u + v) + p(u + v) + q = 0.$$

We now demand that $uv = -p/3$, whence it follows that

$$u^3 + v^3 = -q \tag{5.6}$$

Our condition for the uv implies that

$$u^3 v^3 = -\frac{p^3}{3}\,. \tag{5.7}$$

Together, equations (5.6) and (5.7) state that u^3 and v^3 solve the equation

$$y^2 + qy - \left(\frac{p}{3}\right)^3 = 0,$$

which has solutions $y_{1,2} = -\frac{q}{2} \pm \sqrt{-D}$. It is now easy to check that we may choose the cubic roots u_0 and v_0 of y_1 and y_2 so that the condition $u_0 v_0 = -\frac{p}{3}$ is satisfied and that all possible choices are given by the formula (5.5) □

Remarks.

- Let the coefficients p, q be real numbers. Then one may show that

    1. If $D > 0$, there are 3 different real roots.
    2. If $D < 0$, there is 1 real root and 2 complex roots.
    3. If $D = 0$, there are 2 real roots if $p \neq 0$ and only 1 real root if $p = 0$.

- In the case $D > 0$ one needs to take 3rd root of a complex number, and it can be shown that this operation cannot in general be done by formulas that involve only real roots, which is quite fittingly referred to as the ”casus irreducibilis”! However, the roots may be expressed in terms of trigonometric functions, which actually fits well to our next topic.

5.3 Trisecting angles by a ruler and a compass

A very classical problem of Greek mathematics was whether one can trisect a given angle ϕ (i.e., construct an angle $\phi/3$) by a ruler and a compass. It took almost 2000 years to prove that this impossible! We will sketch a proof this fact.

Theorem 5.11. *Starting from given points in the plane, a geometric construction using a ruler and a compass produces points whose coordinates can be expressed by expressions containing a finite number of divisions, additions, multiplications, taking differences and taking square roots applied on the coordinates of the initial points.*

Proof.

- Note first that a line is determined by two given points, a circle by a given center-point and the radius as the distance between the two given points.

- It follows that any constructed line has equation of the form

$$ax + by = c,$$

where a, b and c are rational expressions of coordinates of the already known points.

- Similarly, a circle that can be constructed has the form

$$(x - a)^2 + (y - b)^2 = r^2,$$

where a, b and r are given. In terms of the coordinates of given points and rational operations combined with taking square roots (e.g. $r = \sqrt{(x_1 - x_2)^2 - (y_1 - y_2)^2}$ for two already known points (x_1, y_1) and (x_2, y_2)).

- It remains to show that intersections of circles or lines can be expressed in terms of rational operations and square roots applied on their equations.

- For the intersection of two lines the previous point is clear (why?).

- For the intersection of a line and a circle, we have

$$\begin{cases} ax + by = c \\ (x - u)^2 + (y - v)^2 = r^2, \end{cases}$$

where the case $b = 0$ is easy. Otherwise, we substitute $y = \frac{c-ax}{b}$ to the equation of the circle and obtain a 2nd degree equation for x. The claim therefore follows by recalling the solution formula of the 2nd degree equation.

- For the intersection of two circles, we have

$$\begin{cases} (x - u_1)^2 + (y - v_1)^2 = r_1^2 \\ (x - u_2)^2 + (y - v_2)^2 = r_2^2. \end{cases}$$

We may assume that either $u_1 \neq u_2$ or $v_1 \neq v_2$ (why?). Subtracting the equations side by side yields

$$2(u_2 - u_1)x + 2(v_2 - v_1)y = A,$$

where $A = r_1^2 - r_2^2 - u_1^2 + u_2^2 - v_1^2 + v_2^2$. The rest is concluded as in the previous case.

□

Lemma 5.12. If one may trisect by a ruler and a compass the angle of 60 degrees, then one may construct a solution of the equation

$$y^2 - 3y - 1 = 0$$

by a finite combination of rational operations (the new coefficients one may add must be also rational numbers) and taking square roots starting from rational numbers.

Proof. The points $(0,0)$, $(1,0)$ and $(\cos 60°, \sin 60°) = (\frac{1}{2}, \frac{\sqrt{3}}{2})$ determine an angle of 60 °. If one could trisect 60 °, one could construct the point $(\cos 20°, \sin 20°)$ (think about this!), which would show that $\cos 20°$ is obtained by rational and square root operation from $(0, 1, \frac{1}{2}, \frac{\sqrt{3}}{2})$, or in any case from $\mathbb{Q}$. Hence it remains to note the following identity (Exercise!):

$$\cos 3x = 4\cos^3 x - 3\cos x$$

Hence if $z = \cos 20°$ we have

$$\cos 60° = \frac{1}{2} = 4z^3 - 3z$$

or $8z^3 - 6z - 1 = 0 \equiv y^3 - 3y - 1 = 0$, whence $y = 2z$. □

Let us now apply this knowledge in abstract algebra.

Lemma 5.13. Let $F \subset \mathbb{R}$ be a subfield; that is, $\{0,1\} \subset F$ and F is closed under addition, multiplication and the following operations, when $x \neq 0$: $\to -x$ and $x \to \dfrac{1}{x}$. Then, if $D \in F$ satisfies $D > 0$ and $\sqrt{D} \notin F$, the set

$$F(\sqrt{D}) := \{a + b\sqrt{D} \mid a, b, \in F\} \subset \mathbb{R}$$

is also a subfield of $\mathbb{R}$ and

$$\mathbb{Q} \subset F \subsetneq F(\sqrt{D}) \tag{5.8}$$

Moreover, for $x \in F(\sqrt{D})$, the representation $x = a + b\sqrt{D}$ is unique.

Proof. Since $1 \in F$, it follows that $n = 1 + 1 + ... + 1 \in F$, for any $n \in \mathbb{N}$, when there are n ones in the sum. Furthermore, $-n = 0 - n \in F$ and finally $m \cdot \frac{1}{n} \in F$ for any $m \in \mathbb{Z}$, $n \in \mathbb{N}$, so $\mathbb{Q} \subset F$ ($\mathbb{Q}$ is the minimal subfield subfield of $\mathbb{R}$, as one might guess!).
Clearly $F \subset F(\sqrt{D})$ and the inclusion is strict ("$\subsetneq$") since $\sqrt{D} \in F(\sqrt{D}) \setminus F$. It remains to check that $F(\sqrt{D})$ is closed under the desired operations. This is clear for addition and the operation $x \to -x$. It is true for multiplication as well, since if $a_1, a_2, b_1, b_2 \in F$, we have

$$(a_1 + b_1\sqrt{D})(a_2 + b_2\sqrt{D}) = (a_1a_2 + b_1b_2D) + (a_1b_2 + a_2b_1)\sqrt{D},$$

where both terms of the sum are members of F. Finally, if $a + b\sqrt{D} \in F(\sqrt{D})$ is nonzero, then

$$\frac{1}{a + b\sqrt{D}} = \frac{a - b\sqrt{D}}{a^2 - b^2D} = \left(\frac{a}{a^2 - b^2D}\right) + \left(\frac{-b}{a^2 - b^2D}\right)\sqrt{D},$$

where both terms of the last expression are once again members of F. Uniqueness is proven exactly as Lemma 5.4. □

The following lemma is crucial for understanding the trisection problem.

Lemma 5.14. Assume that F is a subfield of $\mathbb{R}$. Assume that the 3:rd degree equation

$$x^3 + ux^2 + vx + w = 0, \tag{5.9}$$

where $u, v, w \in F$, is such that none of its roots belong to F. Then none of its roots belongs to a field of the form $F(\sqrt{D})$, where $D > 0$ with $D \in F$ and $\sqrt{D} \notin F$.

Proof. Assume $a + b\sqrt{D}$, $a, b \in F$ and $b \neq 0$, solves (5.9). We obtain

$$(a + b\sqrt{D})^3 + u(a + b\sqrt{D})^2 + v(a + b\sqrt{D}) + w = 0,$$

or equivalently,

$$(a^3 + 3ab^2D + ua^2 + ub^2D + va + w) + (3a^2b + b^3D + 2uab + vb)\sqrt{D} = 0.$$

By the uniqueness of the representation, both quantities in the brackets must be zero. Then the same computation (just a change of the sign in front of $\sqrt{D}$) yields that $a - b\sqrt{D}$ also solves the equation 5.5. In our situation these are different non-zero roots. Since the product of all three roots is $-w$, we deduce that one of the roots is

$$\frac{-w}{(a + b\sqrt{D})(a - b\sqrt{D})} = \frac{w}{b^2D - a^2} \in F,$$

which is contradiction, so the claim is proven. □

We are now ready for the proof of the following theorem concerning the trisection problem.

Theorem 5.15. *Trisection of angles by a ruler and a compass is not always possible.*

Proof. By the Lemma 5.12, it is enough to show that the equation

$$y^3 - 3y - 1 = 0 \tag{5.10}$$

is not solvable by finite expressions containing only square roots and rational numbers. Let us assume to the contrary: then one of the roots of equation (5.10), say y_0, belongs to the field F_n, where

$$\begin{aligned} &\mathbb{Q} = F_0 \subsetneq F_1 \subsetneq ... \subsetneq F_n \ni y_0, \qquad F_{k+1} = F_k(\sqrt{D_k}), \\ &\text{and} \quad D_k > 0 \ \text{ with } \ \sqrt{D_k} \notin F_k \quad \text{for} \quad k = 0, 1, ..., n-1. \end{aligned}$$

Moreover, we may assume that n is minimal, i.e. that none of roots of the equation belong to F_{n-1} (if $n \geq 1$, we discuss the possibility $n = 0$ below separately).

Case $n = 0$: Now $y_0 \in F_0 = \mathbb{Q}$, but equation (5.10) has no rational roots; if $y_0 = \dfrac{k}{r}$ solves the equation and $k \in \mathbb{Z}$, $r \in \mathbb{Z} \setminus \{0\}$, we get

$$k^3 - 3kr^2 - r^3 = 0.$$

We may assume $(k,r)=1$. The above equality implies that $k\,|\,r^3$, whence $k=\pm 1$. Analogously, $r=\pm 1$, but the values $y=\pm 1$ do not solve equation (5.10). Hence $n=0$ is impossible.

Case $n \geq 1$: By our assumption, none of the roots of equation (5.10) belong to F_{n-1}. Now Lemma 5.14 states that none of them can belong to F_n, which is a contradiction! □

5.4 On transcendental numbers

The next theorem is called "Liouville's Approximation Theorem" after its discoverer, the French mathematician Joseph Liouville.

Theorem 5.16. *If $x \notin \mathbb{Q}$ is an algebraic number of degree $n \geq 2$, then there is a constant $c > 0$ (that may depend on x) so that*

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

for all $p, q \in \mathbb{Z}$, $q \geq 1$.

Proof. Let $a_n x^n + a_{n-1}x^{n-1} + ... + a_0 = 0$, $n \geq 2$, $a_n \neq 0$, $a_n, ..., a_0 \in \mathbb{Z}$. Let us denote $f(y) = a_n y^n + a_{n-1}y^{n-1} + ... + a_0$. By continuity, there is $M < \infty$ such that

$$|f'(y)| \leq M, \text{ for } |y-x| \leq 1.$$

It is enough to consider rational numbers p/q with

$$\left| \frac{p}{q} - x \right| \leq \min(1, \frac{r}{2}),$$

where r is the distance between x and the nearest other root of f. Then $f(p/q) \neq 0$ as $p/q \neq x \notin \mathbb{Q}$. We obtain

$$|f(p/q)| = \frac{|a_n p^n + a_{n-1}p^{n-1}q + a_0 q^{n-1}|}{|q|^n} \geq \frac{1}{|q|^n}, \tag{5.11}$$

since the numerator is a non-zero integer. By the mean value theorem,

$$f(p/q) = f(p/q) - f(x) = \Big(\frac{p}{q} - x\Big) f'(\xi),$$

where ξ is between $\frac{p}{q}$ and x. By this equality and estimate (5.11) we may deduce

$$\left| \frac{p}{q} - x \right| = \frac{|f(p/q)|}{|f'(\xi)|} \geq \frac{1}{|q|^n \cdot M}.$$

□

Corollary 5.17. For all the choices of the signs the numbers

$$\xi = 1 \pm \frac{1}{2^{1!}} \pm \frac{1}{2^{2!}} \pm \frac{1}{2^{3!}} \pm ...$$

are transcendental. This expression is referred to as the "Liouville's construction".

Proof. Let us fix the signs. Set $q_n = 2^{n-1!}$, $n \geq 1$ and $p_n = q_n(1 \pm 2^{-1!} \pm 2^{-2!} \pm ... \pm 2^{-(n-1)!})$. Then $\frac{p_n}{q_n}$ is a partial sum of the series and

$$\begin{aligned} \left|\xi - \frac{p_n}{q_n}\right| &\leq 2^{-n!} + 2^{-(n+1)!} + ... \\ &\leq 2^{-n!}(1 + \frac{1}{2} + \frac{1}{4} + ...) \leq 2 \cdot 2^{-n!} \\ &= 2q_n^{-n} \end{aligned}$$

As one may check that $\xi \notin \mathbb{Q}$ (the proof of this is left as an Exercise!), the above estimate that is is true for all $n \geq 2$ and Theorem 5.16 imply that ξ cannot be algebraic. □

Remarks.

- One may check that all the above numbers are different (this is left as an Exercise). Hence Liouville's construction produces uncountably many transcendental numbers.
- Liouville's proof for this is from 1844. Cantor proved the existence of transcendental numbers in 1874 by showing that $\mathbb{R}$ is uncountable (we already know that the algebraic numbers form a countable set).
- π and e are concrete examples of transcendental numbers, but the proof of this fact is rather technical, so we omit it.

Chapter 6

Continued fractions

6.1 Introduction to the notion

Let us now apply the Euclidean algorithm on numbers 67 and 24:

$$\begin{aligned}
67 &= 2 \cdot 24 + 19 \\
24 &= 1 \cdot 19 + 5 \\
19 &= 3 \cdot 5 + 4 \\
5 &= 1 \cdot 4 + 1 \\
\Rightarrow \frac{67}{24} &= 2 + \frac{19}{24} \\
\frac{24}{19} &= 1 + \frac{5}{19} \\
\frac{19}{5} &= 3 + \frac{4}{5} \\
\frac{5}{4} &= 1 + \frac{1}{4}.
\end{aligned}$$

Henceforth we have

$$\begin{aligned}
\frac{5}{4} &= 1 + \frac{1}{4} \\
\Rightarrow \frac{19}{5} &= 3 + \cfrac{1}{1 + \cfrac{1}{4}} \\
\Rightarrow \frac{24}{19} &= 1 + \cfrac{1}{3 + \cfrac{1}{1 + \cfrac{1}{4}}} \\
\Rightarrow \frac{67}{24} &= 2 + \cfrac{1}{1 + \cfrac{1}{3 + \cfrac{1}{1 + \cfrac{1}{4}}}}.
\end{aligned}$$

We have thus produced a continued fraction representation for $\frac{67}{24}$!

There are different notations in use for continued fractions: e.g., the following two are common:

$$\frac{67}{24} = 2 + \frac{1}{1+}\frac{1}{3+}\frac{1}{1+}\frac{1}{4} \text{ or}$$
$$\frac{67}{24} = \{2; 1, 3, 1, 4\}$$

In these lectures we shall employ the latter expression above.

One may encounter infinite continued fractions in a natural manner. Let us illustrate with an example.

Example. Let $x = \dfrac{3+\sqrt{13}}{2} > 1$ so that

$$x^2 - 3x - 1 = 0.$$

We may approximate x by noting that

$$x = 3 + \frac{1}{x} = 3 + \frac{3}{3+\frac{1}{x}} = 3 + \frac{3}{3+\frac{1}{3+\frac{1}{x}}} ...$$

This looks promising, and thus raises intuitive questions:

- Does the infinite continued fraction $\{3; 3, 3, 3, ...\}$ converge? If so, how quick is the convergence?
- Can one develop all (irrational) reals in the same manner?

We will soon learn that all answers are positive and, e.g., one has

$$\left| x - 3 + \frac{1}{3+\frac{1}{3+\frac{1}{3}}} \right| < 10^{-4}.$$

We now start the serious study of continued fractions.

Definition 6.1. Let $\lambda_1, \lambda_2, ..., \lambda_n > 0$. Set

$$\lambda_0 + \frac{1}{\lambda_1 + \frac{1}{\lambda_2 + \frac{1}{... + \lambda_n}}} = \{\lambda_0; \lambda_1, \lambda_2, ..., \lambda_n\}.$$

From what we did at the beginning of this section, it is easy to generalize what we did for arbitrary rationals (instead of 67/24) and obtain the following theorem.

Theorem 6.2. *Every rational $\alpha \in \mathbb{Q}$ can be written as a finite continued fraction*

$$\alpha = \{\lambda_0; \lambda_1, \lambda_2, ..., \lambda_n\},$$

where $\lambda_i \in \mathbb{Z}$, $\forall i \geq 0$ and $\lambda_i > 0$ for $i \geq 1$.

Remark. One may show that the continued fraction representation is unique if we assume that the last denominator, λ_n, is greater or equal to 2, when $n \geq 1$ (Exercise).

Example. $2 = 1 + \frac{1}{1} = 2$, $1\frac{4}{3} = 1 + \cfrac{1}{1 + \cfrac{1}{4}} = 1 + \cfrac{1}{1 + \cfrac{1}{3 + \cfrac{1}{1}}}$. One may easily check that these two ways are the only two alternatives for expressing these rational numbers. We do not prove this in detail since our interest is in developing irrational numbers into continued fractions.

Definition 6.3. The continued fraction $\{\lambda_0; \lambda_1, \lambda_2, ..., \lambda_n, ...\}$ is simple if $\lambda_0 \in \mathbb{Z}$ and $\lambda_0, \lambda_1, ..., \lambda_n, ... \in \mathbb{N}$.

Remark. In this course, however, we will only consider simple continued fractions, and hence the word ”simple” is usually omitted.

6.2 Covergence of infinite continued fractions

We now turn our focus on infinite continued fractions.In order to define the continued fraction of a given irrational number α, assume that

$$\alpha := \{\lambda_0; \lambda_1, \lambda_2, ..., \lambda_n, ...\} \in \mathbb{R} \setminus \mathbb{Q}$$

and furthermore that the right-hand side of this converges suitably — we will soon consider these questions in a general! situation! Then we have

$$\alpha = \lambda_0 + \frac{1}{\lambda_1 + ...}$$

Here $\lambda_1 + \frac{1}{\lambda_2 + ...} > 1$, for otherwise $\alpha \in \mathbb{Q}$. Thus$\lambda_1 + \frac{1}{\lambda_2 + ...} \in]0, 1[$ and we obtain

$$\lambda_0 = \lfloor \alpha \rfloor$$

If we denote $\alpha_0 = \alpha$, and

$$\alpha_1 := \frac{1}{\alpha - \lambda_0} = \frac{1}{\alpha - \lfloor \alpha \rfloor} = \lambda_1 + \frac{1}{\lambda_2 + ...}$$

the same reasoning verifies that $\alpha_1 > 1$ and

$$\lambda_1 = \lfloor \alpha_1 \rfloor$$

By iterating, we have $\lambda_2 = \lfloor \alpha_2 \rfloor$ with $\alpha_2 \coloneqq \dfrac{1}{\alpha_1 - \lfloor \alpha_1 \rfloor}$ and $\lambda_2 = \lfloor \alpha_2 \rfloor$ etc.

The above discussion gives rise to the following definition.

Definition 6.4. Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$. Then the "continued fraction development" of α is

$$\{\lambda_0; \lambda_1, \lambda_2, ...\},$$

where $\lambda_k = \lfloor \alpha_k \rfloor$, and α_k is defined by the recursion

$$\begin{cases} \alpha_0 = \alpha \\ \alpha_{k+1} = \dfrac{1}{\alpha_k - \lfloor \alpha_k \rfloor}, \qquad k \geq 0. \end{cases}$$

Example. Let $\alpha = \sqrt{2}$. Then $\alpha_0 = \sqrt{2}$, $\alpha_1 = \dfrac{1}{\sqrt{2}-1} = \sqrt{2}+1$, $\alpha_2 = \dfrac{1}{(\sqrt{2}+1)-2} = \dfrac{1}{\sqrt{2}-1} = \sqrt{2}+1$, and inductively $\alpha_k = \sqrt{2}+1$ for all $k \geq 1$. Hence, we may write (formally at this point)

$$\sqrt{2} = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + ...}}}} = \{1; 2, 2, 2, 2, ...\}.$$

The following result is fundamental.

Theorem 6.5. *Let $\{\lambda_0; \lambda_1, \lambda_2, ...\}$ be an arbitrary infinite (simple) continued fraction. Define the integer sequences p_k and q_k, $k \geq 0$, by setting*

$$\begin{cases} p_0 = \lambda_0 \\ p_1 = \lambda_1\lambda_0 + 1 \end{cases} \qquad \begin{cases} q_0 = 1 \\ q_1 = \lambda_1 \end{cases} \quad \textit{and}$$

$$\begin{cases} p_k = \lambda_k p_{k-1} + p_{k-2} \\ q_k = \lambda_k q_{k-1} + q_{k-2}, k \geq 2. \end{cases}$$

Then for all $n \geq 0$,

i) $\{\lambda_0; \lambda_1, \lambda_2, ..., \lambda_n\} = \dfrac{p_n}{q_n}$*, which is called the "nth convergent"*

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

iii) Set $p_{-1} = 1$ and $q_{-1} = 0$. Then for every $x \in \mathbb{R}$, $x > 0$, one has

$$\{\lambda_0; \lambda_1, \lambda_2, ..., \lambda_n, x\} = \frac{xp_n + p_{n-1}}{xq_n + q_{n-1}}. \tag{6.1}$$

Proof. (iii) Let us first note that by the very definition $q_n \geq 1$ for all $n \geq 0$ (also we have $q_n \to \infty$ as $n \to \infty$ (why?)). By induction on n:
Base case 1, $n = 0$: $\{\lambda_0; x\} = \lambda_0 + \dfrac{1}{x} = \dfrac{x\lambda_0 + 1}{x} = \dfrac{xp_0 + p_{-1}}{xq_0 + q_{-1}}$
Base case 2, $n = 1$: $\{\lambda_0; \lambda_1, x\} = \lambda_0 + \dfrac{1}{\lambda_1 + \dfrac{1}{x}} = \lambda_0 + \dfrac{x}{\lambda_1 x + 1} = \dfrac{(\lambda_0\lambda_1 + 1)x + \lambda_0}{\lambda_1 x + 1} =$
$\dfrac{p_1 x + p_0}{q_1 x + q_0}$.
Induction step: Assume the claim is true for some $n \geq 1$. Then it follows that

$$\begin{aligned} \{\lambda_0; \lambda_1, ..., \lambda_{n+1}, x\} &= \{\lambda_0; \lambda_1, ..., \lambda_n, \lambda_{n+1} + \frac{1}{x}\} \\ &= \frac{p_n(\lambda_{n+1} + 1/x) + p_{n-1}}{q_n(\lambda_{n+1} + 1/x) + q_{n-1}} = \frac{(\lambda_{n+1}p_n + p_{n-1})x + p_n}{(\lambda_{n+1}q_n + q_{n-1}x + q_n} \\ &= \frac{p_{n+1}x + p_n}{q_{n+1}x + q_n}, \end{aligned}$$

which is the claim for $n + 1$.

(i) By using part (iii) we obtain

$$\begin{aligned} \{\lambda_0; \lambda_1, ..., \lambda_n\} &= \lim_{x\to\infty} \{\lambda_0; \lambda_1, ..., \lambda_n + \frac{1}{x}\} \\ &= \lim_{x\to\infty} \{\lambda_0; \lambda_1, ..., \lambda_n, x\} = \lim_{x\to\infty} \frac{xp_n + p_{n-1}}{xq_n + q_{n-1}} \\ &= \frac{p_n}{q_n} \end{aligned}$$

(ii) Easy induction that is based on the recursion formulas of p_n and q_n. Left as an Exercise.

□

Let us record an immediate and useful corollary of Theorem 6.5ii):

Corollary 6.6. The representation $\frac{p_n}{q_n}$ in the Theorem 6.5 for the nth convergent is in reduced form, i.e. $(p_n, q_n) = 1$ for all $n \geq 0$.

Remark. The recursion formulas of p_k and q_k remain true for $k = 1$. Hence one could define p_k and q_k, $k \geq 1$, by giving the initial value p_{-1}, p_0, q_{-1}, q_0 and defining the rest by the given recursion.

In order to prove the convergence of a given infinite continued fraction we employ the classical result from the analysis courses, called "Leibniz's theorem".

Theorem 6.7. *Let $u_1 > u_2 > ... > u_n \to 0$, as $n \to \infty$, Denote $s_n = u_1 - u_2 + ... + (-1)^{n-1}u_n$, $n \geq 1$. Then*

$$0 < s_2 < s_4 < ... < s_{2n} < s_{2n+1} < s_{2n-1} < ... < s_1$$

for all n. Also the series converges. If we denote its sum by $\beta := \lim_{n\to\infty} s_n$, one has

$$|\beta - s_n| \leq u_{n+1} \qquad \text{for all } n \geq 1.$$

Theorem 6.8. *Let $\frac{p_n}{q_n}$ be the nth convergent of an infinite continued fraction. Then*

i) $q_n < q_{n+1}$, and $q_n \geq 2^{\frac{n}{2}-1}$ for all $n \geq 1$.

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

iii) $\frac{p_0}{q_0} < \frac{p_2}{q_2} < ... < \frac{p_{2n}}{q_{2n}} < \frac{p_{2n+1}}{q_{2n+1}} < \frac{p_{2n-1}}{q_{2n-1}} < ... < \frac{p_1}{q_1}$.

iv) $\exists \lim_{n\to\infty} \frac{p_n}{q_n} = \alpha$ and $\left|\alpha - \frac{p_n}{q_n}\right| \leq \frac{1}{q_n^2} \leq 2^{2-n}$, $n \geq 1$.

v) α is irrational.

Proof. i) Clearly (q_n) is increasing by the formula of q_k on the Theorem 6.5. Furthermore, if $n \geq 1$,

$$q_{n+1} = \lambda_n q_n + q_{n-1} > q_n.$$

We then prove by induction that $q_n \geq 2^{\frac{n}{2}-1}$. This is obviously true if $n = 0$ or $n = 1$. Assume that the claim holds up to some n, Then

$$\begin{aligned} q_{n+1} &= \lambda_{n+1} q_n + q_{n-1} \geq q_n + q_{n-1} \\ &\geq 2^{\frac{n}{2}-1} + 2^{\frac{n-1}{2}-1} \geq (\sqrt{2}+1) 2^{\frac{n-1}{2}-1} \\ &\geq 2^{\frac{n+1}{2}-1}, \end{aligned}$$

which is the claim for $n + 1$.

ii) Follows immediately from ii) of Theorem 6.5.

iii) and iv): We apply Leibniz's Theorem on the sequence

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

Clearly by part *i*), u_n, $n \geq 1$, fills the condition of Leibniz's Theorem. Hence

$$s_n = \frac{1}{q_n q_0} + (-1)^1 \cdot \frac{1}{q_2 q_1} + (-1)^{n-1} \frac{1}{q_n q_{n-1}}$$

converges to a limit $\beta \in \mathbb{R}$. We observe that

$$\begin{aligned} s_n &= \left(\frac{p_1}{q_1} - \frac{p_0}{q_0} \right) - \left(\frac{p_1}{q_1} - \frac{p_2}{q_2}) \right) + ... + (-1)^{n-1} \left((-1)^{n-1} \left(\frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}} \right) \right) \\ &= \frac{p_n}{q_n} - \frac{p_0}{q_0} \end{aligned}$$

Thus there exists

$$\lim_{n \to \infty} \frac{p_n}{q_n} = \beta + \frac{p_0}{q_0}. := \alpha$$

Moreover, by the error estimate of Leibniz's Theorem we have

$$\begin{aligned} \left| \frac{p_n}{q_n} - \alpha \right| &= |s_n - \beta| \leq u_{n+1} = \frac{1}{q_{n+1} q_n} \leq \frac{1}{q_n^2} \qquad (6.2) \\ \leq \frac{1}{2^{\frac{n+1}{2}-1}} \frac{1}{2^{\frac{n}{2}-1}} = 2^{\frac{3}{2}-n} < 2^{2-n}. \end{aligned}$$

v) Acording to (6.2) we have $\left| \alpha - \frac{p_n}{q_n} \right| \leq \frac{1}{q_n^2}$ for infinitely many different q_n. At this situation a simple argument (quided Exercise!) verifies that α is irrational.

□

We will next verify that the continued fraction of a given α indeed converges to α! At the same time we obtain a good estimate for the rate of convergence.

Theorem 6.9. *Let $\{\lambda_0; \lambda_1, \lambda_2, ...\}$ be the continued fraction development of a given irrational number $\alpha \in \mathbb{R} \setminus \mathbb{Q}$. Then*

i) $\alpha = \lim\limits_{n\to\infty} \{\lambda_0; \lambda_1, \lambda_2, ..., \lambda_n\}$

ii) $\alpha = \dfrac{p_n\alpha_{n+1} + p_{n-1}}{q_n\alpha_{n++1} + q_{n-1}} = \{\lambda_0; \lambda_1, \lambda_2, ..., \lambda_n, \alpha_{n+1}\}$, $\forall n \geq 0$, *where α_n is as in the Definition* 6.4.

iii) $\dfrac{1}{(q_{n+1} + q_n)q_n} < \left|\alpha - \dfrac{p_n}{q_n}\right| < \dfrac{1}{q_n q_{n+1}} \leq \dfrac{1}{q_n^2}.$

Proof. ii) We first prove by induction that

$$\alpha = \{\lambda_0; \lambda_1, \lambda_2, ..., \lambda_n, \alpha_{n+1}\}, \qquad n \geq 0.$$

Base case: Recall that $\alpha_0 = \alpha$, $\alpha_{n+1} = \dfrac{1}{\alpha_n - \lfloor \alpha_n \rfloor}$, $n \geq 0$ and $\lambda_n = \lfloor \alpha_n \rfloor$. For $n = 0$, the claim is equivalent to

$$\alpha = \{\lambda_0; \alpha_1\} = \lambda_0 + \frac{1}{\alpha_1},$$

which is just the definition of α_1.

Induction step: Assume that the claim is true for n. Then we have

$$\begin{aligned} \alpha_{n+2} &= \frac{1}{\alpha_{n+1} - \lambda_{n+1}} \quad \text{or} \\ \alpha_{n+1} &= \lambda_{n+1} + \frac{1}{\alpha_{n+2}}. \end{aligned}$$

Substituting the latter into the induction hypothesis yields

$$\begin{aligned} \alpha &= \{\lambda_0; \lambda_1, \lambda_2, ..., \lambda_n, \lambda_{n+1} + \frac{1}{\alpha_{n+2}}\} \\ &= \{\lambda_0; \lambda_1, \lambda_2, ..., \lambda_{n+1}, \alpha_{n+2}\}, \end{aligned}$$

which is the claim for the value $n + 1$. Finally, (6.1) gives

$$\{\lambda_0; \lambda_1, \lambda_2, ..., \lambda_n, \alpha_{n+1}\} = \frac{p_n, \alpha_{n+1} + p_{n-1}}{q_n\alpha_{n+1} + q_{n-1}},$$

which completes the proof of ii).

iii) We obtain an exact formula for the error:

$$\begin{aligned} \left|\alpha - \frac{p_n}{q_n}\right| &= \left|\frac{p_n\alpha_{n+1} + p_{n-1}}{q_n\alpha_{n+1} + q_{n-1}} - \frac{p_n}{q_n}\right| \\ &= \frac{|q_n p_{n-1} - q_{n-1}p_n|}{|q_n(q_n\alpha_{n+1} + q_{n+1})|} \\ &= \frac{1}{q_n(q_n\alpha_{n+1} + q_{n-1})} \end{aligned}$$

Here $1 \leq \lambda_{n+1} = \lfloor \alpha_{n+1} \rfloor < \alpha_{n+1} < \lambda_{n+1} + 1$, so that

$$q_n\alpha_{n+1} + q_{n-1} > q_n\lambda_{n+1} + q_{n-1} = q_{n+1} \text{ and}$$
$$q_n\alpha_{n+1} + q_{n-1} < q_n(\lambda_{n+1} + 1) + q_{n-1} = q_{n+1} + q_n.$$

Thus we obtain

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

as was to be shown.

i) This is an immediate consequence of iii) since we know that $q_n \geq 2^{\frac{n}{2}-1}$ by i), and thus

$$\left| \alpha - \frac{p_n}{q_n} \right| < \frac{1}{2^{\frac{n}{2}-1} 2^{\frac{n+1}{2}-1}} < 2^{2-n} \to 0, \text{ as } n \to \infty.$$

□

Corollary 6.10. There is a one-one correspondence between infinite continued fractions and irrational numbers.

Proof. Denote by $(\lambda_0; \lambda_1, \lambda_2, ..., \lambda_n, ...)$ a continued fraction and by $\{\lambda_0; \lambda_1, \lambda_2, ..., \lambda_n, ...\}$ its value. We claim that the mapping from all continued fraction developments to irrational numbers, defined as,

$$\Psi \colon \{\lambda_0; \lambda_1, \lambda_2, ..., \lambda_n, ...\} \to \mathbb{R} \setminus \mathbb{Q},$$
$$\Psi((\lambda_0; \lambda_1, \lambda_2, ..., \lambda_n, ...)) := \{\lambda_0; \lambda_1, \lambda_2, ..., \lambda_n, ...\}$$

is a bijection. It is well-defined by Theorem 6.8. It is a surjection by Theorem 6.9. Finally, since the value $\{\lambda_0; \lambda_1, \lambda_2, ..., \lambda_n, ...\} \in \mathbb{R} \setminus \mathbb{Q}$, it determines the λ_j like in the reasoning on p. 66, and this verifies that Ψ is injective. □

6.3 Continued fractions and Pell's equation

Theorem 6.11. *Assume that $(p, q) = 1$, $q \geq 2$ and $\alpha \in \mathbb{R} \setminus \mathbb{Q}$. If*

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

then $\frac{p}{q}$ is a convergent of the continued fraction decomposition of α.

Proof. Write by the Euclidean algorithm (recall Theorem 6.2)

$$\frac{p}{q} = \{\lambda_0; \lambda_1, \lambda_2, ..., \lambda_n\} = \frac{p_n}{q_n}$$

where $\lambda_0 \in \mathbb{Z}$, $\lambda_1, \lambda_2, ..., \lambda_n \in \mathbb{Z}_+$, $\lambda_n \geq 2$ and p_k, q_k are defined as usual for $k \in \{1, 2, .., n\}$. Clearly, $\dfrac{p}{q} = \dfrac{p_n}{q_n}$ is the nth convergent of any irrational number of the form

$$(\lambda_0; \lambda_1, \lambda_2, ..., \lambda_{n-1}, \lambda_n + x), x \in (0, 1) \setminus \mathbb{Q},$$

and since we also have $\dfrac{p}{q} = (\lambda_0, \lambda_1, \lambda_2, ..., \lambda_{n-1}, \lambda_n - 1, 1)$, analogously $\dfrac{p}{q}$ is the (n + 1):th convergent of any irrational number of the form

$$\begin{aligned}&(\lambda_0; \lambda_1, \lambda_2, ..., \lambda_{n-1}, \lambda_n - 1, 1 + x)\\ =&(\lambda_0; \lambda_1, \lambda_2, ..., \lambda_{n-1}, \lambda_n - 1 + \frac{1}{1+x}), x \in (0, 1) \setminus \mathbb{Q}\end{aligned}$$

Here $-1 + \dfrac{1}{1+x} = \dfrac{-x}{1+x}$ takes all irrational values from $(-\dfrac{1}{2}, 0)$, so put together we have:

$$\begin{aligned}&\frac{p}{q} \quad \text{is a convergent of any irrational of the form}\\ &(\lambda_0; \lambda_1, \lambda_2, ..., \lambda_{n-1}, \lambda_n + x), \quad x \in (-\frac{1}{2}, 1) \setminus \mathbb{Q}.\end{aligned}$$

By continuity, the claim we want to prove follows as soon as we check that the numbers $\dfrac{p}{q} \pm \dfrac{1}{2q^2}$ are contained in the open interval with the endpoints $(\lambda_0; \lambda_1, \lambda_2, ..., \lambda_{n-1}, \lambda_n + a)$ where $a \in \{-\dfrac{1}{2}, 1\}$. By iii) of Theorem 6.5 we have

$$\begin{aligned}&(\lambda_0; \lambda_1, \lambda_2, ..., \lambda_{n-1}, \lambda_n + a) - \frac{p}{q}\\ &= (\lambda_0; \lambda_1, \lambda_2, ..., \lambda_{n-1}, \lambda_n + a) - (\lambda_0; \lambda_1, \lambda_2, ..., \lambda_n)\\ &= \frac{(\lambda_n + a)p_{n-1} + p_{n-2}}{(\lambda_n + a)q_{n-1} + q_{n-2}} - \frac{\lambda_n p_{n-1} + p_{n-2}}{\lambda_n q_{n-1} + q_{n-2}}.\end{aligned}$$

We may compute this by using Theorem 6.5ii) and obtain the value

$$= \frac{(-1)^{n+1}a}{((\lambda_n + a)q_{n-1} + q_{n-2})(\lambda_n q_{n-1} + q_{n-2})}.$$

By noting that $\lambda_n q_{n-1} + q_{n-2} = q_n = q$, we finally obtain

$$(\lambda_0; \lambda_1, \lambda_2, ..., \lambda_{n-1}, \lambda_n + x\} - \frac{p}{q} = \frac{(-1)^{n+1}a}{(q + aq_{n-1})q}.$$

It remains to note that

$$\begin{aligned}\left|\frac{-(-1)^{n+1}\frac{1}{2}}{(q - \frac{1}{2}q_{n-1})q}\right| = \frac{1}{2q(q - \frac{1}{2}q_{n+1})} \geq \frac{1}{2q^2} \text{ and}\\ \left|\frac{(-1)^{n+1}1}{(q + q_{n+1})q}\right| = \frac{1}{(q + q_{n-1})q} \geq \frac{1}{2q^2}.\end{aligned}$$

□

Theorem 6.12. *Let $D \geq 2$ and $\sqrt{D} \notin \mathbb{Q}$. Then every positive solution to the Pell's equation*

$$x^2 - Dy^2 = 1$$

is given by a convergent of $\sqrt{D}$, i.e., $(x, y) = (p_n q_n)$ for some $n \geq 1$.

Proof. Assume that $x, y \geq 1$ and $x^2 - Dy^2 = 1$. Then $\dfrac{x}{y} = \sqrt{D + \dfrac{1}{y^2}} \geq \sqrt{D}$. Hence

$$\begin{aligned} \left|\frac{x}{y} - \sqrt{D}\right| &= \frac{1}{y}\left|x - y\sqrt{D}\right| = \frac{|x^2 - Dy^2|}{y(x + y\sqrt{D})} \\ &= \frac{1}{y^2(\frac{x}{y} + \sqrt{D})} \leq \frac{1}{2\sqrt{D}y^2} < \frac{1}{2y^2}. \end{aligned}$$

The claim follows now from Theorem 6.11. □

The previous Theorem yields a practical and effective algorithm to find the smallest positive solutions (x_1, y_1) to a given Pell's equation. One simply computes the continued fraction of $\sqrt{D}$ and substitutes the convergents to the equation; the first convergent that solves $x^2 - Dy^2 = 1$ gives the fundamental solution. Let us take a look at an example.

Example. Solve the equation $x^2 - 21y^2 = 1$.

Solution.

$$\begin{aligned} \sqrt{21} &= 4 + \sqrt{21} - 4, \\ \alpha_1 = \frac{1}{\sqrt{21} - 4} &= \frac{4 + \sqrt{21}}{5} = 1 + \frac{\sqrt{21} - 1}{5}, \\ \alpha_2 = \frac{5}{\sqrt{21} - 1} &= \frac{1 + \sqrt{21}}{4} = 1 + \frac{\sqrt{21} - 3}{4}, \\ \alpha_3 = \frac{4}{\sqrt{21} - 3} &= \frac{\sqrt{21} + 3}{3} = 2 + \frac{\sqrt{21} - 3}{3}, \\ \alpha_4 = \frac{3}{\sqrt{21} - 3} &= \frac{\sqrt{21} + 3}{4} = 1 + \frac{\sqrt{21} - 1}{4}, \\ \alpha_5 = \frac{4}{\sqrt{21} - 3} &= \frac{\sqrt{21} + 1}{5} = 1 + \frac{\sqrt{21} - 4}{5}, \\ \alpha_6 = \frac{5}{\sqrt{21} - 4} &= \sqrt{21} + 4 = 8 + \sqrt{21} - 4, \\ \alpha_7 = \frac{1}{\sqrt{21} - 4} &= \alpha_1 \in (0, 1). \end{aligned}$$

Hence the continued fraction is <u>periodic</u>:

$$\sqrt{21} = \{4; \overline{1, 1, 2, 1, 1, 8}, \overline{1, 1, 2, 1, 1, 8}, ...\}$$
$$\lambda_0 = 4, \lambda_1 = \lambda_2 = 1, \lambda_3 = 2, \lambda_4 = \lambda_5 = 1, \lambda_6 = 8, ...$$

We shall compute the convergents by the recursion formulas

$$\begin{cases} p_n = \lambda_n p_{n-1} + p_{n-2} \\ q_n = \lambda_n q_{n-1} + q_{n-2} \end{cases}$$

with $p_0 = \lambda_0 = 4, p_1 = \lambda_0\lambda_1 + 1 = 5, q_0 = 1$ and $q_1 = \lambda_1 = 1$. We thus have

k	p_k	q_k	$p_k^2 - 21q_k^2$
0	4	1	-5
1	5	1	4
2	9	2	-3
3	23	5	4
4	32	7	-5
5	55	12	1
$\cdots$	$\cdots$	$\cdots$	$\cdots$

Hence the fundamental solution is $(x_1, y_1) = (55, 12)$.

We finish by some remarks on the the error term:

- By Theorem 6.9 the error term is of the order

$$\left|\frac{p_n}{q_n} - \alpha\right| \geq \frac{1}{q_{n+1}q_n} \text{ and } \left|\frac{p_n}{q_n} - \alpha\right| \leq \frac{1}{2q_{n+1}q_n}$$

  Hence the upper bound $\frac{1}{q_{n+1}q_n}$ is very sharp and gives the right order.

- Especially, if the λ_{n+1} is unusually large, one may expect that $\{\lambda_0; \lambda_1\lambda_2, ..., \lambda_n\}$ gives a spectacularly good approximation to α. An example of this is given by π: the direct computation that uses a good decimal approximation of π shows that

$$\pi = \{3; 7, 15, 1, 292, ...\}$$

  Here $\{3; 7, 15, 1\} = 3 + \cfrac{1}{7 + \cfrac{1}{15+1}} = \frac{355}{113}$, and $(\pi - \frac{355}{113}) \leq 3 \cdot 10^{-7}$, which is the "Ludolf's number", known already in China in the 4th century.

- Every convergent is a very good approximation of α by Theorem 6.9. Conversely, by Theorem 6.8, every "good enough" approximation is convergent. One can actually get an "if and only if" characterization of convergents as so-called best approximations: $\frac{p}{q}$ is a convergent of α if and only if it satisfies

$$|s - r\alpha| > |p - q\alpha|,$$

  for all other pairs (s, r) with $1 \leq r \leq q$. We do not prove this result in this course.

6.4 Periodic continued fractions

Recall that every rational number has a periodic decimal expansion. For algebraic second order numbers there is an analogue of this in terms of the continued fractions (see the Theorem 6.14 below). We shall start with a helpful lemma.

Lemma 6.13. i) Let $A, A', ..., D, D' \in \mathbb{Z}$ with $C, C' \neq 0$ and $AC' - CA' \neq 0$. If x satisfies

$$\frac{Ax+B}{Cx+D} = \frac{A'x+B'}{C'x+D'},$$

then x is an algebraic number of degree at most 2.

ii) Let $y = \dfrac{Ax+B}{Cx+D}$ with $AD - BC \neq 0$ and $A, B, C, D \in \mathbb{Z}$. Then, if x is an algebraic number of degree ≤ 2, then y is an algebraic number of degree ≤ 2 as well.

Proof. i) Simply multiply out and note that x satisfies an equation with integer coefficients and with $AC' - CA'$ as the coefficient of x^2.

ii) Left as an Exercise for the reader.

□

The next theorem is known as ”Lagrange’s Theorem”.

Theorem 6.14. *The continued fraction of an irrational number α is periodic if and only if α is an algebraic number of degree two.*

Proof. Let $\alpha = \{\lambda_0; \lambda_1, \lambda_2, ...\}$ and assume first that $\{\lambda_0; \lambda_1, \lambda_2, ...\}$ is periodic, i.e., there exists an index $l \geq 1$ so that

$$\lambda_{n+l} = \lambda_n \text{ for } n \geq n_0, \quad \text{with (say) } n_0 \geq 2. \tag{6.3}$$

Define $\alpha_n, n \geq 0$, as before. By this definition (or by the *ii*) of Theorem 6.9), we have that

$$\alpha = \{\lambda_0; \lambda_1, \lambda_2, ..., \lambda_{n-1}, \alpha_n\}$$
$$\alpha_n = \{\lambda_n; \lambda_{n+1}, \lambda_{n+2}, ...\}, \text{ for } n \geq 0.$$

Especially, $\alpha_{n+l} = \alpha_n$ for $n \geq 0$. By the *ii*) of Theorem 6.9, we obtain

$$\alpha = \frac{p_{n_0-1}\alpha_{n_0} + p_{n_0-2}}{q_{n_0-1}\alpha_{n_0} + q_{n_0-2}} = \frac{p_{n_0+l-1}\alpha_{n_0+l} + p_{n_0+l-2}}{q_{n_=+l-1}\alpha_{n_0+l}q_{n_0+l-2}}$$
$$= \frac{p_{n_0+l-1}\alpha_{n_0} + p_{n_0+l-2}}{q_{n_0+l-1}\alpha_{n_0} + q_{n_0+l-2}}$$

The equality of the second and fourth term and the fact $\frac{p_{n_0-1}}{q_{n_0-1}} \neq \frac{p_{n_0+l-1}}{q_{n_0+l-1}}$ imply in view of Lemma 6.13ii) that α_{n_0} is an algebraic number of degree two or less. In turn, part

ii) of the same Lemma and the first equality above $(\frac{p_{n_0-1}}{q_{n_0-1}} \neq \frac{p_{n_0-2}}{q_{n_0-2}})$ imply that α is an algebraic number of degree at most two. Since α is irrational, the degree must be two.

In order to prove the other direction we assume that

$$A\alpha^2 + B\alpha + C = 0,$$

where $A, B, C \in \mathbb{Z}$ and $A \neq 0$. We let $k \geq 2$ and substitute $\alpha = \frac{p_{k-1}\alpha_k + p_{k-2}}{q_{k-1}\alpha_k + q_{k-2}}$ into this equation, getting

$$A(p_{k-1}\alpha_k + p_{k-2})^2 + B(p_{k-1}\alpha_k + p_{k-2})(q_{k-1}\alpha_k + q_{k-2}) + C(q_{k-1}\alpha_k + q_{k-2})^2 = 0,$$

or, equivalently,

$$A_k\alpha_k^2 + B_k\alpha_k + C_k = 0, \qquad \text{where} \tag{6.4}$$

$$\begin{cases} A_k := Ap_{k-1}^2 + Bp_{k-1}q_{k-1} + Cq_{k-1}^2 \neq 0 \\ B_k := 2Ap_{k-1}p_{k-1} + B(p_{k-1}q_{k-2} + p_{k-2}q_{k-1}) + 2Cq_{k-1}q_{k-2} \\ C_k := Ap_{k-2}^2 + Bp_{k-2}q_{k-2} + Cq_{k-2}^2 \end{cases}$$

Write $f(x) := Ax^2 + Bx + C$, so that

$$A_k = q_{k-1}^2 f(\frac{p_{k-1}}{q_{k-1}}) \text{ and } C_k = A_{k-1}.$$

We know that $f(\alpha) = 0$ and $\left|\alpha - \frac{p_{k-1}}{q_{k-1}}\right| \leq \frac{1}{q_{k-1}^2}$ by part ii) of Theorem 6.9. Hence, the mean value theorem yields for all $k \geq 2$:

$$A_k \leq q_{k-1}^2 \left| f(\frac{p_{k-1}}{q_{k-1}}) - \alpha \right| = q_{k-1}^2 \left| f'(\xi) \right| \left| \alpha - \frac{p_{k-1}}{q_{k-1}} \right| \ \leq \ q_{k-1}^2 M q_{k-1}^{-2} \ \leq \ M,$$

where ξ is between α and $\frac{p_{k-1}}{q_{k-1}}$, and $M = \sup_{|s-\alpha|\leq 1} f'(\xi)$. Thus we have shown that the sequences A_k and C_k, $k \geq 2$ are bounded! Furthermore B_k is bounded since there is the identity

$$B_k^2 - 4A_kC_k = B^2 - 4AC, \forall k \geq 2,$$

which follows by direct computation and the fact that

$$p_{k-1}q_{k-2} - p_{k-2}q_{k-1} = \pm 1.$$

We leave the details of this as a computational exercise.

By the proven boundedness there are only finitely many different triples (A_n, B_n, C_n). We may thus pick three indexes $n_1 < n_2 < n_3$ so that

$$(A_{n_1}, B_{n_1}, C_{n_1}) = (A_{n_2}, B_{n_2}, C_{n_2}) = (A_{n_3}, B_{n_3}, C_{n_3})$$

As the second degree equation (6.4) has only two roots, we must have either $\alpha_{n_1} = \alpha_{n_2}$, or $\alpha_{n_1} = \alpha_{n_3}$ or $\alpha_{n_2} = \alpha_{n_3}$. In any case we have found n_0 and $l_0 \geq 1$ so that

$$\alpha_{n_0+l} = \alpha_{n_0},$$

and hence the continued fraction is indeed periodic by the iterative definition of the sequence $(\alpha_k)_{k\geq 1}$ □

Chapter 7

Diophantine equations and Gaussian integers

- Diophantine equations derive their name from Diophantus's book Arithmetica written in 3rd century AD. Only six out of 13 books have survived, and in these are equations to be solved in integers (or rationals).
- Due to Fermat (who spoke of Diophantus's work praisingly in his letters to his peers) a new level of research emerged. Fermat conjectured his famous (or infamous?) Last Theorem as a comment to Diophantus's problem. This question became one of the most studied questions in number theory for hundreds of years, and it was proven true by Andrew Wiles only in 1996.
- We have already solved several Diophantine equations:linear equations, the Pell's equation and the general second degree equation
  $$ax^2 + bx + c = my.$$
- Our main goal in this section is to decide when bi-quadratic equation
  $$x^2 + y^2 = m$$
  is solvable. For that purpose we develop the basic theory of Gaussian integers. Before that, however, we take a look at a couple of simpler examples of Diophantine equations.

Example from *Arithmetica.* Let $p \in \mathbb{Q}$. Determine all rationals x, y such that

$$\frac{x^2 + y^2}{x + y} = p.$$

Solution (*by Diophantus*). Denote $\frac{y}{x} = t \in \mathbb{Q}$ (if $x \neq 0$). Then we get

$$\frac{x^2(1+t^2)}{x(1+t)} = p \Rightarrow \begin{cases} x = \dfrac{p(1+t)}{1+t^2} \\ y = \dfrac{pt(1+t)}{1+t^2} \end{cases}$$

A direct substitution shows that all these solve the given equation. In addition $(x, y) = (0, p)$ is a solution.

7.1 Pythagorean triples

Pythagorean triples are three positive numbers (x, y, z) such that they satisfy the equation

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

Theorem 7.1. *All Pythagorean triples are given by the formulas*

$$(x, y, z) = t(a^2 - b^2, 2ab, a^2 + b^2),$$

or (interchanging x and y)

$$(x, y, z) = t(2ab, a^2 - b^2, a^2 + b^2),$$

where $t \geq 1$, $a > b \geq 1$.

Proof. A direct computation verifies that the triples given by the formula are Pythagorean. In order to prove the converse, we may divide a common divisor out so that $gcd(x, y, z) = 1$. Then (why?)

$$1 = gcd(x, y) = gcd(y, z) = gcd(z, x).$$

Especially, both x and y cannot be even. Both of them cannot be odd either, since otherwise

$$z^2 \equiv 1 + 1 \equiv 2 \pmod 4,$$

which is impossible. Thus one of x and y is even, and the other is odd. Assume, for instance, that $2 \,|\, y$. Now $y^2 = (z - x)(z + x)$. Since z and x are odd, we may write $z - x = 2u$ and $z + x = 2v$, so that

$$x = v - u, \;\; z = v + u.$$

We have $gcd(v, u) = 1$, as otherwise $gcd(x, z) > 1$. As

$$\left(\frac{y}{2}\right)^2 = uv, \;\; gcd(u, v) = 1,$$

it follows that both u and v must be squares (Exercise!). Denote $u = a^2$ and $v = b^2$. Then $y = 2ab$, $z = a^2 + b^2$ and $x = b^2 - a^2$, as was to be shown. □

Example. The choices $t = 1$ and $(a, b) = 1$ or $(a, b) = (3, 2)$ produce the following Pythagorean triples

$$3^2 + 4^2 = 5^2 \text{ and } 5^2 + 12^2 = 13^2.$$

7.2 Gaussian integers

Carl Friedrich Gauss made a fundamental discovery 1800s by noticing that adjoining of $i = \sqrt{-1}$ to integers produces a ring that is very useful for many arithmetic purposes.

We assume that the reader knows the complex numbers $\mathbb{C}$ and their basic properties. In general if $x+yi \in \mathbb{C}$ ($i=\sqrt{-1}$, $x,y \in \mathbb{R}$), we denote

$$\overline{x+iy} = x - iy, \text{ "complex conjugate"}$$
$$|z| = \sqrt{x^2+y^2} = \sqrt{z\overline{z}}.$$

Recall that $\dfrac{1}{z} = \dfrac{\overline{z}}{z\overline{z}} = \dfrac{x-iy}{x^2+y^2}$ etc.

Definition 7.2. The complex number $a+ib \in \mathbb{C}$ is a *Gaussian rational* if $a,b \in \mathbb{Q}$. The set of all Gaussian rationals is denoted by $\mathbb{Q}[i]$.

Remark. Clearly $\mathbb{Q}[i]$ is a subfield of $\mathbb{C}$, for $0,1 \in \mathbb{Q}[i]$ and if $z_1, z_2 \in \mathbb{Q}[i]$, then $-z_1 \in \mathbb{Q}[i]$ and $\dfrac{1}{z_1} \in \mathbb{Q}[i]$ (if $z_1 \neq 0$). In addition, $z_1+z_2 \in \mathbb{Q}[i]$ and $z_1z_2 \in \mathbb{Q}[i]$.

Definition 7.3. The Gaussian integers $\mathbb{Z}[i]$ is the subring of $\mathbb{Q}[i]$, consisting of numbers

$$\mathbb{Z}[i] := \{a+ib \mid a,b \in \mathbb{Z}\}.$$

Recall that an element λ of a ring is a unit if also λ^{-1} belong to the ring.

Theorem 7.4. $\lambda \in \mathbb{Z}[i]$ *is a unit if and only if* $\lambda \in \{\pm 1, \pm i\}$.

Proof. Let $\lambda = a+ib \neq 0$, where $\lambda \in \mathbb{Z}[i]$ so that $a,b \in \mathbb{Z}$. Then

$$\frac{1}{\lambda} = \frac{1}{a+ib} = \frac{a}{a^2+b^2} - \frac{ib}{a^2+b^2}$$

This belongs to $\mathbb{Z}[i]$ if $\dfrac{a}{a^2+b^2} \in \mathbb{Z}$ and $\dfrac{b}{a^2+b^2} \in \mathbb{Z}$. If both $a \neq 0$ and $b \neq 0$, then $|a| < a^2+b^2$, which is impossible. If $a=0$, we must have $\dfrac{1}{b} \in \mathbb{Z}$ whence $b = \pm 1$. Similarly, if $b=0$, we get $a = \pm 1$. □

Definition 7.5. If $a+ib \in \mathbb{Z}[i]$ (or, more generally, $a+ib \in \mathbb{Q}[i]$), the "norm" of $a+ib$ is defined by the formula

$$N(a+ib) = a^2+b^2.$$

Remark. $N(\lambda) = \lambda\overline{\lambda} = |\lambda|^2$, which equals the square of the Euclidean norm.

Theorem 7.6. *i)* $N(\lambda_1\lambda_2) = N(\lambda_1)N(\lambda_2)$

ii) $N(\lambda) = 0 \quad \iff \quad \lambda = 0.$

iii) If $\lambda \in \mathbb{Z}[i]$*, then* $\quad N(\lambda) = 1 \quad \iff \quad \lambda$ *is a unit.*

Proof. The proof is left as an Exercise. □

Definition 7.7. Let $\lambda, \mu \in \mathbb{Z}[i]$. We say that λ divides μ (i.e. λ is a divisor of μ) if $\mu = k\lambda$, where $k \in \mathbb{Z}[i]$. This is denoted by $\lambda \,|\, \mu$.

Theorem 1.2 (see page 4) remains valid in $\mathbb{Z}[i]$ with the same proof.

Definition 7.8. The "associates" of $\lambda \in \mathbb{Z}[i]$ are $\pm\lambda$ and $\pm i\lambda$. In other words, μ is an associate of λ is $\dfrac{\mu}{\lambda}$ is a unit.

Example. $-2+i$ and $1+2i$ are associates since $-2+i = i(1+2i)$.

Definition 7.9. $\lambda \in \mathbb{Z}[i]$ is a prime (in $\mathbb{Z}[i]$) if λ is not a unit and its only divisors are 1, λ and their associates.

The first part of the following theorem provides examples of Gaussian primes (it is not a characterisation though!).

Theorem 7.10. *i) If* $N(\lambda) = p \in \mathbb{P}$*, then* λ *is a prime in* $\mathbb{Z}[i]$*.*

ii) If $\lambda \in \mathbb{Z}[i]$*,* $\lambda \neq 0$ *and* λ *is not a unit, then* λ *is a product of primes.*

Proof. i) If $\lambda = \lambda_1\lambda_2 \Rightarrow N(\lambda_1)N(\lambda_2) = p \in \mathbb{P}$, then $N(\lambda_1) = 1$ or $N(\lambda_2) = 1$ so that either λ_1 or λ_2 is a unit.

ii) Let $\lambda \in \mathbb{Z}[i]$ be a minimal norm element in the set of Gaussian integers which are not products of primes. Then $\lambda = \lambda_1\lambda_2$, where λ_1 and λ_2 are not units, as λ is not a prime itself. Thus $N(\lambda_1), N(\lambda_2) > 1$ and $N(\lambda_1)N(\lambda_2) = N(\lambda)$. Especially, then $N(\lambda_1), N\lambda_2) < N(\lambda)$, which implies that λ_1, λ_2 are hence products of primes, making λ to have the same property. The contradiction proves the claim.

□

Example. $2+i$ is a Gaussian prime since $N(2+i) = 2^2+1 = 5$ is a prime.

Theorem 7.11. *$\mathbb{Z}[i]$ is a Euclidean integral domain, i.e., if $\lambda, \mu \in \mathbb{Z}[i]$ and $\lambda \neq 0$, thee are $k, r \in \mathbb{Z}[i]$ so that*

$$\mu = k\lambda + r, \ \textit{ and } N(r) < N(\lambda).$$

Proof. Denote $\nu = \dfrac{\mu}{\lambda} \in \mathbb{Q}[i] \subset \mathbb{C}$ and write $\nu = a + ib$, $a, b \in \mathbb{Q}$.
If $\nu \in \mathbb{Z}[i]$, we may choose $k = \nu$, $r = 0$. Otherwise we pick $k_1, k_2 \in \mathbb{Z}$ so that

$$|k_1 - a| \leq \frac{1}{2} \text{ and } |k_2 - b| \leq \frac{1}{2}.$$

Set $k := k_1 + k_2 i \in \mathbb{Z}[i]$ and note that

$$N(k - \nu) \leq \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 \leq \frac{1}{2}.$$

Write $r := \mu - k\lambda$ so that $\mu = k\lambda + r$. We may compute

$$N(r) = N(\mu - k\lambda) = N(\lambda(\nu - k)) = N(\lambda)N(\nu - k) \leq \frac{1}{2}N(\lambda) < N(\lambda).$$

□

Remark. Theorem 7.11 is a version of remainder theorem for $\mathbb{Z}[i]$ (for the remainder theorem, see Theorem 1.7 on the page 5).

Lemma 7.12. If $\lambda_1, \lambda_2 \in \mathbb{Z}[i]$ are nonzero and $\lambda_1 \mid \lambda_2$ as well as $\lambda_2 \mid \lambda_1$, the λ_1 and λ_2 are associates.

Proof. The proof is left as an Exercise for the reader. □

Definition 7.13. If $\lambda_1, \lambda_2, \ldots, \lambda_n \in \mathbb{Z}[i]$, we set

$$I(\lambda_1, \lambda_2, \ldots, \lambda_n) = \{\mu_1\lambda_1 + \mu_2\lambda_2 + \cdots + \mu_n\lambda_n \mid \mu_1, \mu_2, \ldots, \mu_n \in \mathbb{Z}[i]\},$$

which is the "ideal" generated by $\lambda_1, \ldots, \lambda_N$.

Theorem 7.14. *For all $\lambda_1, \lambda_2, \ldots, \lambda_n \in \mathbb{Z}[i]$ there is $d \in \mathbb{Z}[i]$ such that*

$$I(\lambda_1, \lambda_2, \ldots, \lambda_n) = I(d) = \{\mu d \mid \mu \in \mathbb{Z}[i]\}.$$

Proof. We may assume that at least one of $\lambda_j \neq 0$ as otherwise we could choose $d = 0$. We choose d from $I(d)$ so that $d \neq 0$ and $N(d)$ is minimal so that especially $N(d) > 0$. Clearly

$$I(d) \subset I(\lambda_1, \lambda_2, \ldots, \lambda_n) \tag{7.1}$$

If the converse inclusion would not hold, there would be

$$h \in I(\lambda_1, \lambda_2, \ldots, \lambda_n) \setminus I(d)$$

By the Euclidean property of $\mathbb{Z}[i]$ (Theorem 7.11), we may write

$$h = kd + r,\ \ k, r \in \mathbb{Z}[i], r \neq 0,$$

with $N(r) < N(d)$. Then $r = h - kd \in I(\lambda_1, \lambda_2, \ldots, \lambda_n)$ and $0 < N(r) < N(d)$, which contradicts the choice of d. Hence there must be equality in inclusion (7.1). □

According to the following corollary the Gaussian integer d in Theorem 7.14 satisfies the properties of the greatest common divisor.

Corollary 7.15. If $\lambda_1, \lambda_2, \ldots, \lambda_n \in \mathbb{Z}[i]$ are not all zero, there exists $d \in \mathbb{Z}[i]$ such that

i) $d \,|\, \lambda_j$, $\forall j = 1, 2, \ldots, n$.

ii) If $d' \,|\, \lambda_j \forall j = 1, 2, \ldots, n$, then $d' \,|\, d$.

We write $d \in gcd(\lambda_1, \lambda_2, \ldots, \lambda_n)$. All greatest common divisors of $\lambda_j, j \in \{1, 2, \ldots, n\}$ (i.e., numbers that satisfy *i)* and *ii)*) are associates.

Proof. Let d be as in the previous Theorem, i.e., $I(d) = I(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Clearly $d \,|\, \lambda_j, 1 \leq j \leq n$. Since $d = \mu_1\lambda_1 + \mu_2\lambda_2 + \cdots + \mu_n\lambda_n$ for some $\mu_1, \mu_2, \ldots \mu_n \in \mathbb{Z}[i]$, any common divisor of $\lambda_1, \lambda_2, \ldots, \lambda_n$ divides d. Hence d satisfies *i)* and *ii)* of the Theorem 7.14.
If d' is another greatest common divisor of $\lambda_1, \lambda_2, \ldots, \lambda_n$, then $d \,|\, d'$ and $d' \,|\, d$, whence d and d' are associates of each other. Conversely, any associate of d is clearly the greatest common divisor f $\lambda_1, \lambda_2, \ldots, \lambda_n$ as well. □

Remarks.

- If $gcd(\lambda_1, \lambda_2, \ldots, \lambda_n)$ consists only of units we denote $gcd(\lambda_1, \lambda_2, \ldots, \lambda_n) = 1$.
- If $gcd(\lambda_1, \lambda_2) = 1$, we say that λ_1 and λ_2 are relatively prime (or mutually prime, co-prime or that they do not have common factors).

Theorem 7.16. *If $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{Z}[i]$, $gcd(\lambda_1, \lambda_2) = 1$ and $\lambda_1 \,|\, \lambda_2\lambda_3$, then $\lambda_1 \,|\, \lambda_3$.*

Proof. Now $1 = \mu_1\lambda_1 + \mu_2\lambda_2$ for some mu_1, μ_2. Thus

$$\lambda_3 = \mu_1\lambda_1\lambda_2 + \mu_2\lambda_2\lambda_3,$$

which implies $\lambda_1 \,|\, \lambda_3$. □

Remarks.

- The proofs above were essentially the same that we gave in Chapter 1 for natural numbers!
- Similarly, the analogues of Corollaries 1.10 and 1.13 (see pages 6 and 7) for Gaussian integers follow from Theorem 7.16 with the same proof as for the standard integers! One just needs to note that the analogue of Corollary 1.13 takes the following form:

Corollary 7.17. If $\lambda \,|\, \mu_1, \mu_2, \ldots, \mu_n$ and $\lambda, \mu_1, \mu_2, \ldots, \mu_n$ are all Gaussian primes, then λ is an associate of one of μ_j, $1 \le j \le n$.

Theorem 7.18. *In $\mathbb{Z}[i]$, every number has a prime decomposition that is unique up to the order of the factors and moving to associate numbers.*

Proof. Again the proof is exactly the same as in the case of regular $\mathbb{Z}$, this time by using Corollary 7.17 instead. □

Remark. We proved the fundamental result given by Theorem 7.18 by using just the results whose proof was based on Theorem 7.11. Hence all Euclidean number fields (i.e., which possess an integer-valued norm satisfying Theorem 7.11) possess a unique prime number decomposition.

Example. Let us try to find the Gaussian prime decomposition of number $5 \in \mathbb{Z}[i]$. Let

$$5 = \lambda_1\lambda_2, \;\; \lambda_1, \lambda_2 \in \mathbb{Z}[i],$$

where $\lambda_1 = a_1 + b_1 i$ and $\lambda_2 = a_2 + b_2 i$. Now

$$25 = N(\lambda_1)N(\lambda_2).$$

IF $N(\lambda_1) = 1$, then λ_1 is a unit and λ_2 is an associate of 5. The situation is symmetric if $N(\lambda_2) = 1$. If instead $N(\lambda_1) = 5$, then

$$a_1^2 + b_1^2 = 5$$

and we may, e.g., choose $a_1 = 2$, $b_1 = 1$ so that $\lambda_1 = 2+i$. Then $\lambda_2 = \dfrac{5}{2+i} = 2-i \in \mathbb{Z}[i]$ and both λ_1, λ_2 are Gaussian primes by the Theorem 7.16 (see also the Example on p.81). Thus

$$5 = (2+i)(2-i),$$

or equivalently, $5 = (-2-i)(-2+i) = (1+2i)(1-2i) = (-1+2i)(-1-2i)$.

Remark. Above, $2+i$ and $2-i$ are different primes since they are not associates:

$$\{\pm(2+i), \pm i(2+i)\} = \{2+i, -2-i, -1+2i, 1-2i\}$$

Example. The four numbers $\pm 1 \pm i$ are primes by the Theorem 7.10, but they are the same in the sense that they are associates of each other.

Example. Is $2-3i$ a factor of $7+i$?

Solution. $\dfrac{7+i}{2-3i} = \dfrac{(7+i)(2+3i)}{2^2+3^2} = \dfrac{11}{13} + \dfrac{23}{13} \notin \mathbb{Z}[i]$. Hence the answer is negative.

Example. Find <u>all</u> the factors of $\lambda = 7+i$.

Solution. We do this here in a reliable but rather cumbersome way; this could later be done more elegantly using the results that we shall soon prove.
If $\mu \,|\, \lambda \Rightarrow N(\mu)|N(\lambda) = 50$. Thus

$$N(\mu) \in \{1, 2, 5, 10, 25, 50\}$$

Trial and error shows that μ must belong to the set

$$\mu \in \{\pm 1, \pm i, \pm 1 \pm i, \pm 1 \pm 2i, \pm 2 \pm i, \pm 3 \pm i, \pm 1 \pm 3i, \pm 4 \pm 3i,\\ \pm 3 \pm 4i, \pm 5, \pm 5i, \pm 5 \pm 5i, \pm 7 \pm i, \pm 1 \pm 7i\}$$

To check, we note that if $\mu = x + iy$, then

$$\begin{aligned} &\frac{7+i}{\mu} \in \mathbb{Z}[i] \\ \iff\quad & \frac{7+i}{x+iy} = \frac{(7+i)(x-iy)}{x^2+y^2} = \frac{(7x+y)+i(x-7y)}{x^2+y^2} \in \mathbb{Z}[i] \\ \iff\quad & \frac{7x+y}{x^2+y^2} \in \mathbb{Z} \text{ and } \frac{x-7y}{x^2+y^2} \in \mathbb{Z} \end{aligned}$$

Now a simple but mechanical checking yields the divisors

$$\{\pm 1, \pm i, \pm 1 \pm i, \pm(2+i), \pm(1-2i), \pm(3-i), \pm(1+3i), \pm(4-3i), \pm(3+4i), \pm(7+i), \pm(1-7i)\}.$$

We then set as our next main goal to determine all the primes in $\mathbb{Z}[i]$. Let us start with a couple of auxiliary results.

Theorem 7.19. *If $\pi \in \mathbb{Z}[i]$ is a prime in $\mathbb{Z}[i]$, then there is a unique standard prime $p \in \mathbb{P}$ (thus $p \in \mathbb{Z}$) such that $\pi \,|\, p$.*

Proof. Now $N(\pi) = \pi\overline{\pi}$, $|\pi| \geq 2$. By writing this as a product of ordinary primes we see that $\pi \,|\, p$ for some $p \in \mathbb{P}$ (the analogue of the Corollary 1.10 is used in this). If there is another $q \in \mathbb{P}$ with $\pi \,|\, q$, we could find $n, n' \in \mathbb{Z}$ so that $nq + n'p = 1$, which would imply $\pi \,|\, 1$ and this is impossible. □

The only slightly more "technical" result from earlier chapters we need for our purposes is the following lemma.

Lemma 7.20. i) If $p \in \mathbb{P}$, $p = 4n + 1$ or $p = 2$, there exists $x \in \mathbb{Z}$ so that

$$p \,|\, x^2 + 1.$$

ii) If $p \in \mathbb{P}$, $p = 4n + 3$, then $p \neq x^2 + y^2$ for all $x, y \in \mathbb{Z}$.

Proof.

i) $\left(\dfrac{-1}{p}\right) = 1$ if $p = 4n + 1 \in \mathbb{P}$ (see the Corollary 4.5 on the page 39).

ii) One checks that $x^2 \not\equiv 1 \pmod 4$ or $x^2 \not\equiv 0 \pmod 4$ for all integers x (enough to consider e.g. $x = 0, \pm 1, 2$). Thus always $x^2 + y^2 \not\equiv a \pmod 4$, where $a \neq 3$.

□

The following theorem is the basis for determining the structure of primes in $\mathbb{Z}[i]$.

Theorem 7.21. *Let $p \in \mathbb{P}$. Then,*

i) If $p = 2$, the prime decomposition of p can be written as

$$2 = (1 + i)(1 - i) = -i(1 + i)^2.$$

Here $1 + i$ is a prime in $\mathbb{Z}[i]$, and $1 - i$ is one of its associates.

ii) If $p = 4n + 1$, then there are $a, b \in \mathbb{Z}$ such that

$$p = (a + ib)(a - ib),$$

and $a \pm ib$ are different primes (not associates) in $\mathbb{Z}[i]$.

iii) If $p = 4n - 1$, then p is a prime in also $\mathbb{Z}[i]$.

Proof.

i) $2 = (1 + i)(1 - i)$, and previously we already noted that $1 \pm i$ are primes which are the essentially the same prime since they are associates of each other.

ii) By Lemma 7.20 there is $x \in \mathbb{Z}$ so that $p \,|\, x^2 + 1$ or $p \,|\, (x - i)(x + i)$. If p would be a prime in $\mathbb{Z}[i]$ it would follow that $p \,|\, (x - i)$ or $p \,|\, (x + i)$, which is impossible since

$$\frac{x \pm i}{p} = \frac{x}{p} \pm \frac{1}{p} i \notin \mathbb{Z}[i].$$

Hence $p = \pi\mu$ where $\pi \in \mathbb{Z}[i]$ is a Gaussian prime and μ is not a unit. Since

$$p^2 = N(p) = N(\pi)N(\mu),\ 1 < N(\pi) < p^2,$$

we must have $|\pi| = p$ so that

$$\pi\overline{\pi} = p \Rightarrow \mu = \overline{\pi},$$

making μ also be a prime (why?). If we write $\pi = a + ib$ with $a, b \in \mathbb{Z}$, we obtain

$$p = (a + ib)(a - ib) = a^2 + b^2. \tag{7.2}$$

We still need to check that $a + ib$ and $a - ib$ are not associates. By (7.2) we have $|a| \neq |b|$ (why?), and none of the associates $\{\pm\pi, \pm i\pi\} = (-a - ib, -b + ia, b - ia)$ can equal $\overline{\pi} = a - ib$.

iii) If p wasn't a prime in $\mathbb{Z}[i]$, we could write $p = \pi\mu$, where π is a prime and μ is not a unit, and deduce as in part *ii*) that

$$p = a^2 + b^2,$$

which is impossible by the *ii*) of Lemma 7.20.

□

Corollary 7.22. The primes in $\mathbb{Z}[i]$ are

i) $1 + i$

ii) $p = 4n - 1 \in \mathbb{P}$,

iii) $a \pm ib$, where $a^2 + b^2 = 4n + 1 \in \mathbb{P}$ and

iv) all the associates of the above numbers.

Proof. If $\pi \in \mathbb{Z}[i]$ is a Gaussian prime, then by Theorem 7.19 it divides an ordinary prime $p \in \mathbb{P}$. The claim now follows from Theorem 7.21. □

Theorem 7.23. *A prime $p \in \mathbb{P}$ can be expressed as a sum of two squares*

$$p = x^2 + y^2$$

if and only if $p = 2$ or p is of the form $p = 4n + 1$. The representation is unique up to the order and signs of x and y.

Proof. The case $p = 2$ is clear. In turn the case where p is of the form $p = 4n - 1$ follows directly from Lemma lemma: ZIprimediv. In the case where $p = 4n + 1$, it is a sum of two squares by part ii) of Theorem 7.21, since it says that

$$p = (a + ib)(a - ib) = a^2 + b^2,$$

where $\pi := a + ib$ and $\overline{\pi} = a - ib$ are Gaussian primes. If $p = (c + id)(c - id) = c^2 + d^2$ is another representation, write

$$\pi' := c + id,$$

whence $N(\pi') = p$ so that π' and $\overline{\pi'}$ are also Gaussian primes. From $\pi\overline{\pi} = \pi\overline{\pi'}$ we deduce that either $\pi' = \varepsilon\pi$ or $\overline{\pi'} = \varepsilon\pi$, with ε being a unit, and this gives the stated uniqueness. □

Example. $7 \equiv 3 \pmod 4$ is not a sum of two squares, but $113 \equiv 1 \pmod 4$ is:

$$113 = 7^2 + 8^2.$$

Lemma 7.24. *$n \in \mathbb{N}$ is a sum fo two squares exactly when $n = \lambda\overline{\lambda} = |\lambda|$ for some $\lambda \in \mathbb{Z}[i]$.*

Proof. $n = x^2 + y^2 \equiv n = (x + iy)(x - iy)$. □

Theorem 7.25. *Let $n \in \mathbb{N}$. Then*

$$n = x^2 + y^2 \text{ with } x, y \in \mathbb{Z},$$

if and only if every prime of the form $4k-1$ that divides n, occurs as an even power in the prime decomposition of n.

Remark. The condition says that if $q \in \mathbb{P}$, $q \equiv -1 \pmod 4$, then if q^β is the highest power of q dividing n, then β is even.

Proof. Let us write the standard prime decomposition of n as

$$n = 2^e p_1^{\alpha_1} p_2^{\alpha_2} \dots p_k^{\alpha_k} q_1^{\beta_1} q_2^{\beta_2} \dots q_l^{\beta_l} \tag{7.3}$$

where p_j and q_j are different primes, $\alpha_1, \alpha_2, \dots, \alpha_k, \beta_1, \beta_2, \dots, \beta_l \geq 1$, $e \geq 0$, $k, l \geq 0$, and $p_j \equiv 1 \pmod 4$, $j = 1, 2, \dots, k$, $q_s \equiv -1 \pmod 4$, $s = 1, 2, \dots, l$.

1. Case $l = 0$ or if $l \geq 1$, every β_s is even.

   By Lemma 7.20 (if $k \geq 1$), we may write $p_j = \lambda_j\overline{\lambda_j}$ for each $j \in \{1, 2, \dots, k\}$, where $\lambda_j \in \mathbb{Z}[i]$. Since now β_s is even for every s, we have

$$\lambda := (1+i)^e \lambda_1^{\alpha_1} \lambda_2^{\alpha_2} \dots \lambda_k^{\alpha_k} q_1^{\frac{\beta_1}{2}} q_2^{\frac{\beta_2}{2}} q_l^{\frac{\beta_l}{2}} \in \mathbb{Z}[i]$$

   and since

$$\begin{aligned} \lambda\overline{\lambda} &= (1+i)^e(1-i)^e \lambda_1^{\alpha_1}\overline{\lambda_1}^{\alpha_1} \dots \lambda_k^{\alpha_1}\overline{\lambda_k}^{\alpha_1} q_1^{\beta_1} \dots q_l^{\beta_l} \\ &= 2^e p_1^{\alpha_1} p_2^{\alpha_2} \dots p_l^{\alpha_l} q_1^{\beta_1} q_2^{\beta_2} \dots q_l^{\beta_l} = p, \end{aligned}$$

   the prime p is a sum of two squares by the Lemma 7.24.

2. Assume that one the β_s is odd.

   We may assume that β_1 is odd. If p is a sum of two squares, we could write

   $$p = \lambda\overline{\lambda},\ \lambda \in \mathbb{Z}[i].$$

   By the Theorem 7.21, q_1 is a Gaussian prime. Especially, it is not a conjugate of any of the λ_i, $\overline{\lambda_i}$ or $q_2, q_3, \ldots, q_l$. If q_1^A is the highest power of q_1 that divides λ, then q_1^A is also the highest power that divides $\overline{\lambda}$ (why?). Thus q_1^{2A} is the highest power that divides p, which is a contradiction since $2A$ is even.

□

The previous result is far from non-trivial! By using he same techniques we can actually determine the number of possible different representations of given n as a sum of two squares.

Definition 7.26. Let $n \in \mathbb{N}$. The number of different ways to express n as a sum of two squares is denoted by

$$r_2(n) := \#\Big(\{(x, y) \in \mathbb{Z}^2 \mid x^2 + y^2 = n\}\Big).$$

Example. $r_2(5) = 8$, since

$$5 = (\pm 1)^2 + (\pm 2)^2 = (\pm 2)^2 + (\pm 1)^2.$$

Similarly, $r_2(4) = 4$, since

$$4 = (\pm 2)^2 + 0^2 = 0^2 + (\pm 2)^2.$$

Theorem 7.27. *Let $n \geq 1$, and assume that*

$$n = 2^{e_0} \prod_{j=1}^{k} p_j^{\alpha_j} \prod_{s=1}^{l} q_s^{\beta_s},$$

where $p_1, p_2, \ldots, p_k, q_1, q_1, \ldots, q_s \in \mathbb{P}$ are different primes with $p_j \equiv 1 \pmod 4$, $q_s \equiv -1 \pmod 4$ for $1 \leq j \leq k$ and $1 \leq s \leq l$. Then $r_2(n) = 0$ if any of the exponents β_s is odd. Otherwise

$$r_2(n) = 4 \prod_{j=1}^{k} (1 + \alpha_j)$$

Proof. By Theorem 7.25 we may assume that all β_s are even. Now we have to count the number of Gaussian integers λ such that

$$\lambda\overline{\lambda} = n = (1 - i)^{e_0}(1 + i)^{e_0} \prod_{j=1}^{k} \lambda_j^{\alpha_j} \prod_{j=1}^{k} \overline{\lambda_j}^{\alpha_j} \prod_{s=1}^{l} q_s^{\beta_s} \tag{7.4}$$

Above, we wrote $p_j = \lambda_j \overline{\lambda_j}$ for a suitable Gaussian prime λ_j, $1 \leq j \leq k$, as in the previous Theorem. We know that $1-i$ is an associate of the Gaussian prime $1+i$, and the set $\{\lambda_1, \overline{\lambda_1}, \lambda_2, \overline{\lambda_2}, \dots, \lambda_k, \overline{\lambda_k}, q_1, q_2, \dots, q_l\}$ consists of different Gaussian primes, which are not associates. By equation (7.4) all the Gaussian prime factors of λ are included in the primes we just mentioned (i.e., prime factors of n). Thus, we may write λ uniquely in the form

$$\lambda = i^v (1+i)^v \prod_{j=1}^{k} \lambda_j^{t_j} \prod_{j=1}^{k} \overline{\lambda_j}^{t'_j} \prod_{s=1}^{l} q_s^{\mu_s},$$

where $v \in \{0,1,2,3\}$ (i^v then gives all different units), with all the exponent $u, t_j, t'_j, \mu_s \geq 0$. Then equation (7.4) holds if

$$\begin{aligned} n = \lambda\overline{\lambda} &= i^v i^{-v} (1+i)^u (1-i)^u \prod_{j=1}^{k} \lambda_j^{t_j} \prod_{j=1}^{k} \overline{\lambda_j}^{t_j} \cdot \prod_{j=1}^{k} \overline{\lambda_j}^{t'_j} \prod_{j=1}^{k} \lambda_j^{t'_j} \prod_{u=1}^{l} q_u^{2\mu_s} \\ &= 2^u \prod_{j=1}^{k} p_j^{t_j + t'_j} \prod_{j=1}^{l} q_k^{2\mu_S} \end{aligned}$$

This holds true assuming that for all j, s is true

$$\begin{cases} u = e_0 \\ t_j + t'_j = \alpha_j \\ 2\mu_s = \beta_s \end{cases}$$

The second equation gives $1 + \alpha_j$ alternatives for each j. Furthermore, there are four different values for v. Put together, there are

$$4 \prod_{j=1}^{k} (1 + \alpha_j)$$

choices for λ. □

Example. If $n = 500$, we may write

$$500 = 2^2 \cdot 5^3.$$

Thus there are $4 \cdot (3+1) = 16$ different ways to express 500 as sums of two squares. We may find the different choices by trial and error, or use the previous proof to write

$$\lambda = i^v (1+i)^2 (2+i)^a (2-i)^{3-a}, 0 \leq a, v \leq 3.$$

Small computation gives

$$\begin{aligned} 500 &= (\pm 4)^2 + (\pm 22)^2 = (\pm 22)^2 + (\pm 4)^2 \\ &= (\pm 10)^2 + (\pm 20)^2 = (\pm 20)^2 + (\pm 10)^2. \end{aligned}$$

7.3 On quadratic number fields

One may also look at other quadratic number fields instead of $\mathbb{Z}[i]$, such as $\mathbb{Q}[\sqrt{D}]$, $D \in \mathbb{Z}$, $\sqrt{D} \notin \mathbb{Z}$. The integers in the field $\mathbb{Q}[\sqrt{D}]$ are, as it turns out, of the form

$$a + bw,\ a, b \in \mathbb{Z},\ \text{where}$$
$$w = \begin{cases} \sqrt{D}, \text{ if } D \not\equiv 1 \pmod 4 \\ \dfrac{1+\sqrt{D}}{2}, \text{ if } D \equiv 1 \pmod 4 \end{cases}$$

A fundamental question arises: when is prime factorization unique in $\mathbb{Q}[\sqrt{D}]$? This is not always the case as shown by the following example and exercise:

Example. In $\mathbb{Q}[\sqrt{D}]$, $6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5})$.

Exercise. Show that $2, 3, 1 \pm \sqrt{-5}$ are (different) primes of $\mathbb{Q}[\sqrt{-5}]$.

One may define a norm in $\mathbb{Q}[\sqrt{D}]$ as well. If the equivalent of the Theorem 7.11 holds for this norm, then the field is Euclidean and our proof of the Theorem 7.16 remains valid. This lead the following result:

- In an Euclidean field prime factorization is unique.

So, when is $\mathbb{Q}[\sqrt{D}]$ Euclidean?

- If $D < 0$, $\mathbb{Q}[\sqrt{D}]$ is Euclidean if and only if $D \in \{-1, -2, -3, -7, -11\}$.
- If $D > 0$, $\mathbb{Q}[\sqrt{D}]$ is Euclidean if and only if

  $$D \in \{2, 3, 5, 6, 7, 11, 13, 17, 19, 21, 29, 33, 37, 41, 57, 73\}.$$

  This proof of this latter fact is more difficult; the first version is from 1950's.

Furthermore, for some other D the prime decomposition is unique as well; several examples include $D \in \{-19, -43, -67, -163\}$.

A famous question (due to Gauss) is still open today: are there infinitely many $D > 0$ such that prime factorization is unique in $\mathbb{Q}[\sqrt{D}]$?

Remark. The ideal theory of German mathematicians Ernst Kummer and Richard Dedekind in a sense reinstates the uniqueness of prime factorization to any $\mathbb{Q}[\sqrt{D}]$. All the above questions make sense in $\mathbb{Q}[x_0]$ as well, where x_0 is any algebraic number.

7.4 Lagrange's theorem on four squares

We finish this course by using Fermat's famous method of infinite descent to show that every positive integer is a sum of four squares. The credit for this result is most likely due to Fermat, even though the first published proof is by Lagrange, since some historians believe that Fermat possibly possessed a proof which used his method of infinite descent. The proof we give is due to Euler (essentially), but could be exactly the kind of proof that Fermat likely had.

Example. $11 = 9 + 1 + 1 + 0$, $4 = 1 + 1 + 1 + 1 = 4 + 0 + 0 + 0$, $7 = 4 + 1 + 1 + 1$. As we can see from the case 7, three squares are not enough!

Theorem 7.28. *For every $n \in \mathbb{N}$, there $x, y, z, w \in \mathbb{Z}$ so that*

$$n = x^2 + y^2 + z^2 + w^2.$$

Proof. We shall soon apply the surprising identity

$$\begin{aligned}
&(x_1^2 + x_2^2 + x_3^2 + x_4^2)(y_1^2 + y_2^2 + y_3^2 + y_4^2) \\
=&(x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4)^2 + (x_1y_2 - x_2y_1 + x_3y_4 - x_4y_3)^2 \\
&\qquad + (x_1y_3 - x_3y_1 + x_4y_2 - x_2y_4)^2 + (x_1y_4 - x_4y_1 + x_2y_3 - x_3y_2)^2
\end{aligned}$$

This shows that if n_1 and n_2 both are sums of four squares, then also the product n_1n_2 is! By iterating this we see that it is enough to show that every odd prime $p \in \mathbb{P}$ is a sum of four squares (since $2 = 1^2 + 1^2 + 0^2 + 0^2$). We shall first show in our next lemma that suitable (not too big) multiples of p are sums of four squares.

Lemma 7.29. If $p \geq 3$, $p \in \mathbb{P}$, then there are $x, y \in \mathbb{Z}$ with

$$1 + x^2 + y^2 = mp, \text{ where } m < p.$$

Proof. Consider the numbers $0^2, 1^2, 2^2, \dots, \left(\frac{p-1}{2}\right)^2$. They are mutually non-congruent modulo p (just recall the proof Theorem 4.3). The same consequently holds true for the numbers

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

Altogether there are $p+1$ integers in these sets and we conclude that one element from the first set must be congruent to one in the other set modulo p, whence there are $x, y \in \{0, 1, 2, \dots, \frac{p-1}{2}\}$ with

$$x^2 \equiv 1 - y^2 \pmod{p}$$

or $1 + x^2 + y^2 = mp$, where $m = \dfrac{1 + x^2 + y^2}{p} \leq \dfrac{1 + 2(\frac{p-1}{2})^2}{p} < p$. □

Let us now start in earnest the proof of Theorem 7.28. Let $m \geq 1$ be the smallest positive integer such that

$$x_1^2 + x_2^2 + x_3^2 + x_4^2 = mp \tag{7.5}$$

for some $x_1, x_2, x_3, x_4 \in \mathbb{Z}$.
By the previous Lemma m is well-define and $m < p$. We want to show that $m = 1$. Assume then that $m > 1$. We shall derive a contradiction.

1. Assume that m is even:

   Then $0, 2$ or 4 of the numbers x_1, x_2, x_3, x_4 in the equation 6.4 are odd. By rearranging the terms, we may assume that either

(a) all x_1, x_2, x_3, x_4 are even or

(b) all x_1, x_2, x_3, x_4 are odd or

(c) x_1, x_2 are odd and x_3, x_4 are even.

In all cases we have that $\dfrac{x_1 \pm x_2}{2}$ and $\dfrac{x_3 \pm x_4}{2}$ are integers. Especially, we obtain that

$$\begin{aligned} &\left(\frac{x_1 + x_1}{2}\right)^2 + \left(\frac{x_1 - x_2}{2}\right)^2 + \left(\frac{x_3 + x_4}{2}\right)^2 + \left(\frac{x_3 - x_4}{2}\right)^2 \\ &\qquad = \frac{1}{2}\left(x_1^2 + x_2^2 + x_3^2 + x_4^2\right) = \frac{m}{2}p, \end{aligned}$$

which proves that m wasn't a minimal and hence gives a contradiction.

2. Assume that m is odd:

Then $m \geq 3$. If all x_i in (7.5) are divisible by m, we would obtain $p \,|\, m$, which is impossible as $m < p$. We may thus select absolutely smallest remainders y_1, y_2, y_3, y_4 and numbers b_1, b_2, b_3, b_4 so that

$$y_j = x_j - b_j m,\ |y_j| < \frac{1}{2}m,\ j \in \{1, 2, 3, 4\}$$

where m is odd and at least one of the y_j is non-zero. Then

$$0 < y_1^2 + y_2^2 + y_3^2 + y_4^2 < 4 \cdot \left(\frac{m}{2}\right)^2 < m^2.$$

Since $y_j \equiv x_j \pmod{m}$, we also have

$$y_1^2 + y_2^2 + y_3^2 + y_4^2 \equiv x_1^2 + x_2^2 + x_3^2 + x_4^2 \equiv 0 \pmod{m}.$$

Thus we obtain

$$\begin{cases} x_1^2 + x_2^2 + x_3^2 + x_4^2 = mp,\ 1 < m < p \\ y_1^2 + y_2^2 + y_3^2 + y_4^2 = m_1 m,\ 1 < m_1 < m \end{cases}$$

By multiplying these equations side by side and applying (7.4) we get

$$\begin{aligned} m^2 m_1 p &= (x_1^2 + x_2^2 + x_3^2 + x_4^2)(y_1^2 + y_2^2 + y_3^2 + y_4^2) \\ &= z_1^2 + z_2^2 + z_3^2 + z_4^2, \text{ where} \\ z_1 &\coloneqq x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4 \\ &\equiv x_1^2 + x_2^2 + x_3^2 + x_4^2 \equiv 0 \pmod{m}, \\ z_2 &\coloneqq x_1 y_2 - x_2 y_1 + x_3 y_4 - x_4 y_3 \\ &\equiv x_1 x_2 - x_2 x_1 + x_3 x_4 - x_4 x_3 \equiv 0 \pmod{m}, \\ z_3 &\coloneqq x_1 y_3 - x_3 y_1 + x_4 y_2 - x_2 y_4 \\ &\equiv x_1 x_3 - x_3 x_1 + x_4 x_2 - x_2 x_4 \equiv 0 \pmod{m}, \\ z_4 &\coloneqq x_1 y_4 - x_4 y_1 + x_2 y_3 - x_3 y_2 \\ &\equiv x_1 x_4 - x_4 x_1 + x_2 x_3 - x_3 x_2 \equiv 0 \pmod{m}, \end{aligned}$$

Thus the numbers z_1, z_2, z_3, z_4 are divisible by m, the numbers $t_j := \frac{z_j}{m}$, $j = 1, 2, 3, 4$, are integers and

$$t_1^2 + t_2^2 + t_3^3 + t_4^2 = m_1 p,$$

where $m_1 < m$, which gives the desired contradiction. Hence proof of the theorem on four squares is complete.

□

Remark. In 1909, David Hilbert solved the famous "Waring's problem" by showing that for any $k \geq 1$, there is a number $r(k) < \infty$ such that every natural number n has the representation

$$n = x_1^k + x_2^k + \cdots + x_{r(k)}^k,\ x_j \geq 0,\ x_j \in \mathbb{Z}.$$