32 Lectures on The Theory of Numbers

Shane Chern

Dalhousie University

E-mail: chenxiaohang92@gmail.com
xh375529@dal.ca
Website: https://shanechern.github.io/

Link: https://shanechern.github.io/publications/lectures/NTv1.pdf

Lecture notes for *MATH 3070 – Theory of Numbers* at Dalhousie University.

Contents

1. Primes

1.1 Divisibility

Definition 1.1 Let a and b be integers. We say that

$$\text{“}a \textit{ divides } b\text{”} \quad \text{or} \quad \text{“}b \textit{ is divisible by } a\text{”}$$

if there exists an integer x such that

$$b = ax.$$

We usually write $a \mid b$ if a divides b. Otherwise, if a does not divide b, we write $a \nmid b$.

Example 1.1 Since $18 = 2 \times 9$, we have $2 \mid 18$; since $35 = 7 \times 5$, we have $7 \mid 35$. ■

Definition 1.2 If $a \mid b$, then a is called a *divisor*, or a *factor*, of b. In particular, a positive divisor of b which is different from b is called a *proper divisor*.

Theorem 1.1 Assume that all variables in this theorem are integers.

(i) $1 \mid a$, $a \mid a$ and $a \mid 0$;
(ii) If $a \mid b$, then $a \mid bc$;
(iii) If $a \mid b$ and $b \mid c$, then $a \mid c$;
(iv) If $a \mid b$, then $ac \mid bc$;
(v) If $a \mid b_i$ for $i = 1, \ldots, r$, then $a \mid (m_1 b_1 + \cdots + m_r b_r)$.

Proof. **(i).** Since $a = 1 \cdot a = a \cdot 1$, we have $1 \mid a$ and $a \mid a$; since $0 = a \cdot 0$, we have $a \mid 0$.

(ii). Note that $a \mid b$ implies that $b = ax$ for a certain integer x. Thus, $bc = (ax) \cdot c = a \cdot (cx)$, implying that $a \mid bc$.

(iii). Note that $a \mid b$ implies that $b = ax$ and that $b \mid c$ implies that $c = by$. Thus, $c = by = (ax) \cdot y = a \cdot (xy)$, implying that $a \mid c$.

(iv). Note that $a \mid b$ implies that $b = ax$. Thus, $bc = (ax) \cdot c = (ac) \cdot x$, implying that $ac \mid bc$.

(v). Note that $a \mid b_i$ implies that $b_i = ax_i$. Thus,

$$m_1 b_1 + \cdots + m_r b_r = \sum_{i=1}^{r} m_i \cdot (ax_i) = a \sum_{i=1}^{r} m_i x_i,$$

implying that $a \mid (m_1 b_1 + \cdots + m_r b_r)$. ■

1.2 Primes

Definition 1.3 A positive integer p is a *prime* if

(i) $p \geq 2$;

(ii) p has no positive divisors other than 1 and p.

A positive integer **greater than** 1 that is not prime is a *composite.*

Example 1.2 The sequence of primes starts with

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

The sequence of composites starts with

$$4, 6, 8, 9, 10, 12, 14, 15, 16, 18, 20, \ldots$$

The number 1 is neither prime nor composite. ■

1.3 Infinitude of primes

Now there is a natural question:

Question 1.1 Does the sequence of primes terminate at some place? Or is it infinite?

The first answer to this question was given over 2,000 years ago by the ancient Greek mathematician Euclid (c. 300 BCE).

Theorem 1.2 (Euclid). The number of primes is infinite.

Proof (of Euclid). Let $\{p_1, \ldots, p_k\}$ be a finite set of primes. Consider

$$n = p_1 p_2 \cdots p_k + 1.$$

Then $n \geq 3$. Note that n has a prime factor p. But p is not one of p_i's; otherwise, we have $p \mid p_1 \cdots p_k$, and since $p \mid n$, it follows that $p \mid (n - p_1 \cdots p_k) = 1$, thereby leading to a contradiction.

Therefore, for any finite set of primes, we are always able to generate a new prime. In other words, a finite set of primes cannot cover all primes. ■

The idea of the above proof is very natural and one may make modifications to establish results in a similar vein.

Theorem 1.3 The number of primes of the form $4s + 3$ is infinite.

Proof. Let $\{p_1, \ldots, p_k\}$ be a finite set of primes. Consider

$$n = 4 p_1 p_2 \cdots p_k - 1.$$

Note that n is of the form $4s + 3$. We claim that n has at least one prime factor p of the form $4s + 3$. Otherwise, if all prime factors of n are of the form $4s + 1$, so is their product, namely, n, thereby leading to a contradiction. Further, the above p is not one of 2, p_1, …, p_k by a similar argument to that for Theorem 1.2. Thus, we arrive at a new prime of the form $4s + 3$ from the set $\{p_1, \ldots, p_k\}$, and hence conclude the infinitude of primes of the form $4s + 3$. ■

Theorem 1.4 The number of primes of the form $6s+5$ is infinite.

Proof. Exercise. ■

R In general, let a and m be positive integers such that $1 \le a \le m$ and $(a,m)=1$. Then the number of primes of the form $ms+a$ is infinite; this is known as *Dirichlet's theorem on primes in arithmetic progressions*, and we will prove it in Sect. 18.3. Furthermore, let $\pi_{a,m}(x)$ count the number of primes not exceeding x that are of the form $ms+a$. For fixed m, let a_1 and a_2 be such that $1 \le a_1, a_2 \le m$ and $(a_1,m)=(a_2,m)=1$. Then

$$\lim_{x\to\infty} \frac{\pi_{a_1,m}(x)}{\pi_{a_2,m}(x)} = 1.$$

1.4 Fermat numbers and the second proof of the infinitude of primes

Definition 1.4 *Fermat numbers* are those of the form $F_n = 2^{2^n}+1$ with $n = 0,1,2,\dots$

On December 25, 1640, the French mathematician Pierre de Fermat wrote to Marin Mersenne:

> *If I can determine the basic reason why*
>
> $$3, 5, 17, 257, 65537, \dots,$$
>
> *are prime numbers, I feel that I would find very interesting results, for I have already found marvelous things [along these lines] which I will tell you about later.*

However, Fermat's conjecture that all F_n are primes was unfortunately proved incorrect as Leonhard Euler discovered in 1732 that

$$F_5 = 4294967297 = 641 \times 6700417.$$

Furthermore, the known prime Fermat numbers, also known as *Fermat primes* are still the five numbers $F_0,\dots,F_4$ examined by Fermat. As of 2014, it is known that F_n is composite for $5 \le n \le 32$. The largest Fermat number currently known to be composite is $F_{18233954}$, and its prime factor $7\times 2^{18233956}+1$ was discovered in October 2020. It is now conjectured that only the first 5 Fermat numbers are prime.

Theorem 1.5 For $n \ge 1$,

$$F_n - 2 = \prod_{i=0}^{n-1} F_i.$$

Proof. We prove this result by induction on n. First, it is true for $n=1$ since $F_1 - 2 = 3 = F_0$. Next, we assume that it is true for $n=k$ with $k \ge 1$. Thus,

$$F_k - 2 = \prod_{i=0}^{k-1} F_i.$$

Now we have

$$F_{k+1} - 2 = (2^{2^{k+1}}+1) - 2 = 2^{2^{k+1}} - 1 = (2^{2^k}+1)(2^{2^k}-1)$$

$$= F_k(F_k - 2) = F_k \cdot \prod_{i=0}^{k-1} F_i$$

$$= \prod_{i=0}^{k} F_i,$$

implying that the statement is also valid for $n = k+1$. ■

Corollary 1.6 Any two distinct Fermat numbers have no common divisor greater than 1.

Proof. Assume that there is a prime p dividing both F_m and F_n with $0 \le m < n$. Since $p \mid F_m$, we have $p \mid \prod_{i=0}^{n-1} F_i$ for p_m appears as a multiplicand of this product. Now, $p \mid F_n$ implies that $p \mid \left(F_n - \prod_{i=0}^{n-1} F_i\right)$, i.e. $p \mid 2$, by Theorem 1.5. Thus, $p = 2$. But this is impossible since all Fermat numbers are odd. ■

Now we are in a position to present the second proof of the infinitude of primes.

Second Proof of Theorem 1.2. Note that the sequence of Fermat numbers is infinite. We collect prime factors of these Fermat numbers, and by Corollary 1.6, they are pairwise distinct. Therefore, there are infinitely many primes. ■

1.5 Fundamental theorem of arithmetic

Theorem 1.7 Every integer $n \ge 2$ is a finite product of primes.

Proof. We prove this by induction on n. First, 2 is a prime itself, and thus the statement is true for $n = 2$. Assume that the statement is true for $n = 2, \ldots, k-1$ with $k \ge 3$. Then if $n = k$ is prime, there is nothing to prove. If $n = k$ is composite, then we may write $k = x \cdot y$ such that $1 < x, y < k$. By our inductive assumption, both x and y are finite products of primes, so is their product $xy = k$. Hence, the statement is true for $n = k$. ■

Now, a natural question is *how many representations are there to factorize $n \ge 2$ as a product of primes?* This question is answered by the *Fundamental Theorem of Arithmetic*, also known as the *Unique Factorization Theorem*.

Fundamental Theorem of Arithmetic Every integer $n \ge 2$ has a unique (up to reordering) representation as a finite product of primes.

This theorem, although intuitionistic, is far more than trivial. We will give its proof in the next lecture.

1.6 Divergence of $\sum_p \frac{1}{p}$ and the third proof of the infinitude of primes

We have a straightforward consequence of the Fundamental Theorem of Arithmetic. Let $n \ge 2$. Consider

$$\prod_{\substack{p \text{ prime} \\ p \le n}} \left(1 + \frac{1}{p} + \frac{1}{p^2} + \cdots\right).$$

If we expand this product, then for each i with all its prime factors no larger than n, the fraction $\frac{1}{i}$ appears as exactly one of the terms in the expansion. In particular, such i's include all integers $m \leq n$. Therefore,

$$\prod_{\substack{p \text{ prime} \\ p \leq n}} \left(1 + \frac{1}{p} + \frac{1}{p^2} + \cdots\right) > \sum_{m=1}^{n} \frac{1}{m}.$$

It follows that

$$\prod_{p \leq n} \frac{1}{1 - \frac{1}{p}} = \prod_{p \leq n} \left(1 + \frac{1}{p} + \frac{1}{p^2} + \cdots\right) > \sum_{m=1}^{n} \frac{1}{m} > \int_1^n \frac{dt}{t} = \log n.$$

On the other hand,

$$\begin{aligned}
\log \prod_{p \leq n} \frac{1}{1 - \frac{1}{p}} &= \sum_{p \leq n} \log \frac{1}{1 - \frac{1}{p}} = \sum_{p \leq n} \sum_{k=1}^{\infty} \frac{1}{k \cdot p^k} \\
&= \sum_{p \leq n} \frac{1}{p} + \sum_{p \leq n} \sum_{k=2}^{\infty} \frac{1}{k \cdot p^k} \\
&< \sum_{p \leq n} \frac{1}{p} + \sum_{p \leq n} \sum_{k=2}^{\infty} \frac{1}{2p^2 \cdot p^{k-2}} \\
&= \sum_{p \leq n} \frac{1}{p} + \sum_{p \leq n} \frac{1}{2p^2} \sum_{k=0}^{\infty} \frac{1}{\cdot p^k} \\
&= \sum_{p \leq n} \frac{1}{p} + \sum_{p \leq n} \frac{1}{2p^2} \frac{p}{p-1} \\
&\leq \sum_{p \leq n} \frac{1}{p} + \frac{1}{2} \sum_{m=2}^{n} \frac{1}{m(m-1)} \\
&< \sum_{p \leq n} \frac{1}{p} + \frac{1}{2}.
\end{aligned}$$

Thus,

$$\sum_{p \leq n} \frac{1}{p} + \frac{1}{2} > \log \prod_{p \leq n} \frac{1}{1 - \frac{1}{p}} > \log \log n.$$

Theorem 1.8 For $n \geq 2$,

$$\sum_{\substack{p \text{ prime} \\ p \leq n}} \frac{1}{p} > \log \log n - \frac{1}{2}. \tag{1.1}$$

In particular, $\sum_{p \text{ prime}} \frac{1}{p}$ diverges.

This result gives the third proof of the infinitude of primes.

Third Proof of Theorem 1.2. If there are finitely many primes, then $\sum_p \frac{1}{p}$ is also finite, thereby contradicting the divergence of $\sum_p \frac{1}{p}$ as established in Theorem 1.8. ■

R In fact, as $x \to \infty$,

$$\sum_{p \leq x} \frac{1}{p} \sim \log \log x,$$

and more precisely,

$$\sum_{p\le x}\frac{1}{p}=\log\log x+M+o(1),$$

where $M\approx 0.2614972128\dots$ is the Meissel–Mertens constant, named after the German astronomer Ernst Meissel and the Polish mathematician Franz Mertens.

1.7 Erdős's proof of the divergence of $\sum_p \frac{1}{p}$

The previous proof of the divergence of $\sum_p \frac{1}{p}$ has, more or less, an analytic flavor. What will be provided here is an elegant elementary attack due to the Hungarian mathematician Paul Erdős (*Mathematica, Zutphen. B.* **7** (1938), 1–2).

Theorem 1.9 The series $\sum_{p \text{ prime}} \frac{1}{p}$ diverges.

Proof. We argue by contradiction. That is, we assume that $\sum_p \frac{1}{p}$ converges. Let $\{p_1,p_2,\dots\}$ be the sequence of primes in increasing order.

First, given an arbitrary positive integer n and an index K, we denote by $N_K(n)$ the number of positive integers $m\le n$ such that the prime factors of m are exclusively from $p_1,\dots,p_K$. Note that by the Fundamental Theorem of Arithmetic, each integer a can be uniquely written as $a=s^2\cdot t$ where t has no square factor other than 1. Meanwhile, the squares no greater than n are 1^2, 2^2, …, $\lfloor\sqrt{n}\rfloor^2$ where $\lfloor x\rfloor$ denotes the largest integer not exceeding a real x. Also, there are 2^K integers of the form $\prod_{i=1}^K p_i^{\varepsilon_i}$ with $\varepsilon_i\in\{0,1\}$. Now, if we write integers m counted by $N_K(n)$ as $m=s^2\cdot t$, then s^2 comes from the above squares and t comes from the above $\prod_{i=1}^K p_i^{\varepsilon_i}$. Hence, $N_K(n)\le 2^K\sqrt{n}$.

On the other hand, the assumption of the convergence of $\sum_p \frac{1}{p}$ means that the index K may be chosen so that $\frac{1}{p_{K+1}}+\frac{1}{p_{K+2}}+\cdots<\frac{1}{2}$. Now we observe that the number $N'_K(n)$ of integers $m'\le n$ with at least one prime factor among $p_{K+1},p_{K+2},\dots$ is bounded by

$$N'_K(n)\le\frac{n}{p_{K+1}}+\frac{n}{p_{K+2}}+\cdots<\frac{n}{2}.$$

Noting that $N_K(n)+N'_K(n)=n$, we obtain that the following holds for any positive integer n:

$$n<2^K\sqrt{n}+\frac{n}{2}.$$

However, it fails when $n=2^{2K+2}$, thereby giving a contradiction. Hence, $\sum_p \frac{1}{p}$ diverges. ■

2. Fundamental theorem of arithmetic

2.1 Greatest common divisor and Euclidean algorithm

Theorem 2.1 Given integers a and b, not both 0. There exists a unique positive integer d such that

(i) $d \mid a$ and $d \mid b$;

(ii) If $\delta \mid a$ and $\delta \mid b$, then $\delta \mid d$.

Definition 2.1 The integer d in Theorem 2.1 is called the *greatest common divisor* of a and b, written as $d = \gcd(a,b) = (a,b)$.

R We may understand (a,b) as the largest positive integer that is a divisor of both a and b.

Definition 2.2 If $(a,b) = 1$, we say that a and b are *relatively prime*, or *coprime*.

The proof of Theorem 2.1 is based on the so-called *Euclidean Algorithm*.

Proof (Euclidean Algorithm). Without loss of generality, we assume that $a \ge b > 0$. We also put $r_{-1} = a$ and $r_0 = b$. Now let us iteratively write

$$r_{-1} = q_1 r_0 + r_1, \qquad 0 < r_1 < r_0; \tag{2.1a}$$

$$r_0 = q_2 r_1 + r_2, \qquad 0 < r_2 < r_1; \tag{2.1b}$$

$$r_1 = q_3 r_2 + r_3, \qquad 0 < r_3 < r_2; \tag{2.1c}$$

$$\cdots$$

$$r_{k-2} = q_k r_{k-1} + r_k, \qquad 0 < r_k < r_{k-1}; \tag{2.1d}$$

$$r_{k-1} = q_{k+1} r_k + 0. \tag{2.1e}$$

We claim that $d = r_k > 0$.

(i). By (2.1e), we have $r_k \mid r_{k-1}$. Then by (2.1d), $r_k \mid r_{k-2}$. Continuing this process, we have $r_k \mid r_0 = b$ and $r_k \mid r_{-1} = a$.

(ii). If $\delta \mid a = r_{-1}$ and $\delta \mid b = r_0$, we know from (2.1a) that $\delta \mid r_1$, and then by (2.1b), $\delta \mid r_2$. Continuing this process, we have $\delta \mid r_k = d$. ■

We may use the Euclidean Algorithm to calculate the greatest common divisor.

■ **Example 2.1** Consider $(1071,462)$:

$$\begin{aligned} 1071 &= 2\times 462+147; \\ 462 &= 3\times 147+\mathbf{21}; \\ 147 &= 7\times 21+0. \end{aligned}$$

Thus, $(1071,462)=21$. ■

Definition 2.3 The greatest common divisor of $n_1,\ldots,n_k$ is the largest positive integer that divides all of $n_1,\ldots,n_k$.

2.2 Modular systems

Definition 2.4 A *modular system* S is a subset of integers such that

(i) If $n\in S$, then $-n\in S$;

(ii) If $m,n\in S$, then $m+n\in S$.

R Modular systems are instances of additive groups under the "+" operation.

■ **Example 2.2** The set of integers $\{\ldots,-2,-1,0,1,2,\ldots\}$ is a modular system. The set of multiples of 3, namely, $\{\ldots,-6,-3,0,3,6,\ldots\}$, is a modular system. Further, the set $\{0\}$ is also a modular system. ■

Theorem 2.2 Let S be a modular system such that $S\neq\emptyset$. Then

(i) $0\in S$;

(ii) If $n\in S$ and x is an integer, then $xn\in S$.

Proof. **(i).** Let $m\in S$ since S is nonempty. Then by definition, $-m\in S$. Finally, $0=m+(-m)\in S$.

(ii). Without loss of generality, we assume that x is a nonnegative integer. Otherwise, we write $xn=(-x)(-n)$. Note that the statement is true for $x=0$ by Part (i). Assume that it is true for $x=0,\ldots,k$ with $k\geq 0$, i.e. $xn\in S$ for $x=0,\ldots,k$. Then for $x=k+1$, we have $(k+1)n=n+kn\in S$ since both n and kn are in S. The statement then follows by induction. ■

Theorem 2.3 Let a and b be integers. Then $S=\{ax+by : x,y\in\mathbb{Z}\}$ is a modular system.

Proof. **(i).** Given any $n\in S$, it is of the form $n=ax+by$ for some integers x and y. Now, $-n=-(ax+by)=a\cdot(-x)+b\cdot(-y)\in S$.

(ii). Given any $m,n\in S$, then they are of the form $m=ax_1+by_1$ and $n=ax_2+by_2$. Now, $m+n=a(x_1+x_2)+b(y_1+y_2)\in S$. ■

Theorem 2.4 Let S be a modular system such that S is neither $\emptyset$ nor $\{0\}$. Let δ be the smallest positive integer in S. Then $S=\{k\delta : k\in\mathbb{Z}\}$.

Proof. We first note that $k\delta\in S$ for all integers k by Theorem 2.2(ii). Now assume that there exists an integer $n\in S$ such that n is not a multiple of δ. Then we may write

$$n=q\delta+r,\qquad 0<r<\delta.$$

This implies that $r = n - q\delta \in S$. But it contradicts the assumption that δ is the smallest positive integer in S. ■

We close this section with a relation named after the French mathematician Étienne Bézout.

Theorem 2.5 (Bézout's Identity). Let a and b be integers, not both 0. Let $d = (a,b)$. Then

$$\{ax + by : x,y \in \mathbb{Z}\} = \{kd : k \in \mathbb{Z}\}.$$

In other words, an integer n can be written as

$$n = ax + by, \qquad x,y \in \mathbb{Z},$$

if and only if n is a multiple of (a,b).

Proof. We write

$$S_1 := \{ax + by : x,y \in \mathbb{Z}\} \qquad \text{and} \qquad S_2 := \{kd : k \in \mathbb{Z}\}.$$

(i). Show $S_1 \subset S_2$. That is, if $n = ax + by$, then $n \in S_2$. This is obvious since both a and b are multiples of $d = (a,b)$, so is $ax + by$.

(ii). Show $S_2 \subset S_1$. That is, there exist integers x and y such that $kd = ax + by$ for any $k \in \mathbb{Z}$. Note that it suffices to prove the case $k = 1$, i.e. $d = ax + by$ or $d \in S_1$. We will require the process in the Euclidean Algorithm. Note that S_1 is a modular system by Theorem 2.3 and $a,b \in S_1$. By (2.1a), $r_1 \in S_1$, and then by (2.1b), $r_2 \in S_1$. Continuing this process, we find that $d = r_k \in S_1$, as desired.

We conclude that $S_1 = S_2$ since they are subsets of one another. ■

2.3 Proof of the fundamental theorem of arithmetic

Let us begin with a crucial implication of Bézout's identity.

Theorem 2.6 If $a \mid bc$ and $(a,b) = 1$, then $a \mid c$.

Proof. By Theorem 2.5, we may find integers x and y such that $1 = ax + by$. Now,

$$c = c \cdot 1 = c \cdot (ax + by) = a \cdot (cx) + (bc) \cdot y.$$

Since bc is a multiple of a, we have $a \mid c$. ■

Corollary 2.7 If a prime $p \mid ab$, then at least one of $p \mid a$ and $p \mid b$ is true.

Proof. If $p \mid a$, then we are done. If $p \nmid a$, then $(p,a) = 1$ since p is a prime. Hence, $p \mid b$ by Theorem 2.6. ■

Corollary 2.8 If a prime $p \mid p_1 p_2 \cdots p_k$ with $p_1, \ldots, p_k$ primes, then $p = p_j$ for at least one j.

Proof. Since $p \mid p_1(p_2 \cdots p_k)$, we have either $p \mid p_1$, which implies $p = p_1$, or $p \mid p_2 \cdots p_k$ by Corollary 2.7. We may then repeat this process for the latter case. ■

Now we are ready to complete the proof of the Fundamental Theorem of Arithmetic.

Theorem 2.9 (Fundamental Theorem of Arithmetic). Every integer $n \geq 2$ has a unique (up to reordering) representation as a finite product of primes.

Proof. In Theorem 1.7, we have shown that every integer $n \geq 2$ is a finite product of primes. It suffices to establish uniqueness. Assume that n has prime factorizations

$$n = p_1 p_2 \cdots p_k = q_1 q_2 \cdots q_\ell.$$

Then $p_1 \mid q_1 q_2 \cdots q_\ell$, and thus by renumbering the q's, we have $p_1 = q_1$ by Corollary 2.8. Dividing by p_1 on both sides, we have

$$p_2 \cdots p_k = q_2 \cdots q_\ell.$$

Repeating this process gives the desired result. ■

Definition 2.5 The *canonical form* of an integer $n \geq 2$ is given by its factorization

$$n = \prod_{j=1}^{k} p_j^{\alpha_j}$$

with p_j the distinct prime factors of n and $\alpha_j > 0$. Also, the canonical form of the integer $n = 1$ is simply $1 = 1$.

Theorem 2.10 If

$$a = \prod_{j=1}^{r} p_j^{\alpha_j} \qquad \text{and} \qquad b = \prod_{j=1}^{r} p_j^{\beta_j},$$

where p_j's are distinct prime factors of either a or b and $\alpha_j, \beta_j \geq 0$, then

$$(a,b) = \prod_{j=1}^{r} p_j^{\min(\alpha_j,\beta_j)}.$$

Proof. We write

$$(a,b) = \prod_{j=1}^{r} p_j^{\delta_j}.$$

Then $\delta_j \leq \alpha_j$ and $\delta_j \leq \beta_j$ but δ_j should not be smaller than both α_j and β_j. ■

2.4 Least common multiple

Definition 2.6 Let a and b be integers with $a, b \neq 0$. Then the *least common multiple* of a and b is the unique positive integer m such that

(i) $a \mid m$ and $b \mid m$;
(ii) If $a \mid \mu$ and $b \mid \mu$, then $m \mid \mu$.

We write $m = \text{lcm}(a,b) = [a,b]$.

R The least common multiple of a and b is the smallest positive integer that is a multiple of both a and b.

Definition 2.7 The least common multiple of $n_1, \ldots, n_k$ is the smallest positive integer that is divisible by all of $n_1, \ldots, n_k$.

Theorem 2.11 If

$$a = \prod_{j=1}^{r} p_j^{\alpha_j} \qquad \text{and} \qquad b = \prod_{j=1}^{r} p_j^{\beta_j},$$

where p_j's are distinct prime factors of either a or b and $\alpha_j, \beta_j \geq 0$, then

$$[a,b] = \prod_{j=1}^{r} p_j^{\max(\alpha_j,\beta_j)}.$$

Proof. This is a direct consequence of the definition of the least common multiple. ■

Theorem 2.12 Let a and b be positive integers. Then

$$[a,b] = \frac{ab}{(a,b)}.$$

Proof. If we write $a = \prod_{j=1}^{r} p_j^{\alpha_j}$ and $b = \prod_{j=1}^{r} p_j^{\beta_j}$, then

$$\begin{aligned}
[a,b]\cdot(a,b) &= \prod_{j=1}^{r} p_j^{\max(\alpha_j,\beta_j)} \cdot \prod_{j=1}^{r} p_j^{\min(\alpha_j,\beta_j)} \\
&= \prod_{j=1}^{r} p_j^{\max(\alpha_j,\beta_j)+\min(\alpha_j,\beta_j)} \\
&= \prod_{j=1}^{r} p_j^{\alpha_j+\beta_j} \\
&= \prod_{j=1}^{r} p_j^{\alpha_j} \cdot \prod_{j=1}^{r} p_j^{\beta_j} \\
&= ab,
\end{aligned}$$

where we make use of the fact that $\max(\alpha,\beta)+\min(\alpha,\beta) = \alpha+\beta$. ■

3. Linear congruences

3.1 Congruences

Definition 3.1 Let m be a positive integer. Let a and b be integers. We say that a *is congruent to* b *modulo* m if

$$m \mid (a-b).$$

We write

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

If $m \nmid (a-b)$, we write

$$a \not\equiv b \pmod{m}.$$

Theorem 3.1 Let m be a positive integer.

(i) $a \equiv a \pmod{m}$;
(ii) If $a \equiv b \pmod{m}$, then $b \equiv a \pmod{m}$;
(iii) If $a \equiv b \pmod{m}$ and $b \equiv c \pmod{m}$, then $a \equiv c \pmod{m}$.

Proof. **(i).** We have $a-a=0$ and $m \mid 0$.

(ii). Since $a \equiv b \pmod{m}$, we have $m \mid (a-b)$. Consequently, m divides $-(a-b)=b-a$, thereby implying that $b \equiv a \pmod{m}$.

(iii). Since $a \equiv b \pmod{m}$ and $b \equiv c \pmod{m}$, we have $m \mid (a-b)$ and $m \mid (b-c)$, and thus $m \mid \big((a-b)+(b-c)\big) = (a-c)$, which yields $a \equiv c \pmod{m}$. ■

R A relation "$\sim$" between the elements of a set is called an *equivalence relation* if it satisfies the conditions:

(i) $a \sim a$ (*reflexivity*);
(ii) If $a \sim b$, then $b \sim a$ (*symmetry*);
(iii) If $a \sim b$ and $b \sim c$, then $a \sim c$ (*transitivity*).

Congruence modulo a fixed positive integer m is an equivalence relation.

Theorem 3.2 We have

(i) $a \equiv b \pmod{m}$ if and only if $a-b \equiv 0 \pmod{m}$;

(ii) If $a_1 \equiv b_1 \pmod{m}$ and $a_2 \equiv b_2 \pmod{m}$, then

$$\begin{aligned} a_1 + a_2 &\equiv b_1 + b_2 \pmod{m}, \\ a_1 a_2 &\equiv b_1 b_2 \pmod{m}; \end{aligned}$$

(iii) If $a \equiv b \pmod{m}$, then for any positive integer k,

$$a^k \equiv b^k \pmod{m};$$

(iv) If $f(x_1, x_2, \ldots)$ is a multivariate polynomial with integer coefficients, and $a_1 \equiv b_1 \pmod{m}$, $a_2 \equiv b_2 \pmod{m}$, ..., then

$$f(a_1, a_2, \ldots) \equiv f(b_1, b_2, \ldots) \pmod{m}.$$

Proof. Exercise. ■

Theorem 3.3 If $a \equiv b \pmod{m}$ and $a \equiv b \pmod{n}$, then

$$a \equiv b \pmod{[m,n]}.$$

R If $(m,n) = 1$, then by Theorem 2.12, we have $[m,n] = \frac{mn}{(m,n)} = mn$. Thus in this case $a \equiv b \pmod{mn}$.

Proof. Since $a \equiv b \pmod{m}$ and $a \equiv b \pmod{n}$, we have $m \mid (a-b)$ and $n \mid (a-b)$. In other words, $a - b$ is a common multiple of m and n, and thus a multiple of $[m,n]$. ■

Note that if $ka \equiv ka' \pmod{m}$, it is *not* always true that $a \equiv a' \pmod{m}$.

■ **Example 3.1** We have $10 \times 1 \equiv 10 \times 4 \pmod{15}$, but $1 \not\equiv 4 \pmod{15}$. However, it is true that $1 \equiv 4 \pmod{3}$ where $3 = \frac{15}{(10,15)} = \frac{15}{5}$. ■

Theorem 3.4 If $(k,m) = d$, then $ka \equiv ka' \pmod{m}$ if and only if $a \equiv a' \pmod{\frac{m}{d}}$.

Proof. We write $k = k_1 d$ and $m = m_1 d$ so that $(k_1, m_1) = 1$. Thus,

$$\frac{ka - ka'}{m} = \frac{k(a-a')}{m} = \frac{k_1(a-a')}{m_1}.$$

Since $(k_1, m_1) = 1$, the left-hand side is an integer if and only if $m_1 \mid (a - a')$, namely, $a \equiv a' \pmod{m_1}$ while we also note that $m_1 = \frac{m}{d}$. ■

Now we can determine on which occasion one may carry out “division” for congruences.

Corollary 3.5 If $(k,m) = 1$, then $ka \equiv ka' \pmod{m}$ if and only if $a \equiv a' \pmod{m}$.

3.2 Residue classes

Definition 3.2 A set $\{a_1, a_2, \ldots, a_m\}$ is called a *complete residue system modulo m*, or a *complete system modulo m*, if

(i) $a_i \not\equiv a_j \pmod{m}$ for any $i \neq j$;

(ii) For any integer a, there exists an index i such that $a \equiv a_i \pmod{m}$.

■ **Example 3.2** **(i).** $\{0,7,2,-3,-8,5\}$ is a complete system modulo 6; **(ii).** $\{0,1,2,\dots,n-1\}$ is a complete system modulo n. ■

R Given a set of m integers, to verify whether it forms a complete system modulo m, it suffices to check if the m integers are pairwise incongruent modulo m.

Theorem 3.6 Let $\{a_1,\dots,a_m\}$ be a complete system modulo m and let k be an integer with $(k,m)=1$. Then $\{ka_1,\dots,ka_m\}$ is also a complete system modulo m.

Proof. **(i).** Show $ka_i \not\equiv ka_j \pmod{m}$ for $i \neq j$. Otherwise, if $ka_i \equiv ka_j \pmod{m}$, then since $(k,m)=1$, we have $a_i \equiv a_j \pmod{m}$ by Corollary 3.5, yielding a contradiction to the assumption that $\{a_1,\dots,a_m\}$ is a complete system modulo m.

(ii). Show $a \equiv ka_i \pmod{m}$ for some i. Since $(k,m)=1$, we may find integers k' and m' such that $kk'+mm'=1$ by Theorem 2.5. It follows that $kk' \equiv 1 \pmod{m}$. Choose the index i such that $a_i \equiv ak' \pmod{m}$. Then $ka_i \equiv k(ak') = a(kk') \equiv a \pmod{m}$. ■

Theorem 3.7 Let m and m' be such that $(m,m')=1$. Suppose that a runs through a complete system modulo m and a' runs through a complete system modulo m'. Then $a'm+am'$ runs through a complete system modulo mm'.

Proof. There are mm' numbers $a'm+am'$. Thus, it suffices to verify that they are pairwise incongruent modulo mm'. Note that if

$$a_1'm+a_1m' \equiv a_2'm+a_2m' \pmod{mm'},$$

then since $(m,m')=1$, it follows from Corollary 3.5 that

$$a_1m' \equiv a_2m' \pmod{m} \qquad \Rightarrow \qquad a_1 \equiv a_2 \pmod{m}$$

and

$$a_1'm \equiv a_2'm \pmod{m'} \qquad \Rightarrow \qquad a_1' \equiv a_2' \pmod{m'}.$$

thereby leading to the same choice of $a'm+am'$ as a runs through a complete system modulo m and a' runs through a complete system modulo m'. ■

3.3 Linear congruences

Theorem 3.8 The linear congruence

$$ax \equiv b \pmod{m} \tag{3.1}$$

is solvable if and only if $(a,m) \mid b$. In this case, there is a unique solution modulo $\frac{m}{(a,m)}$.

Proof. The congruence $ax \equiv b \pmod{m}$ is equivalent to $b-ax=my$ for some y. That is

$$ax+my=b. \tag{3.2}$$

By Theorem 2.5, it has integer solutions (x,y) if and only if b is a multiple of (a,m).

For the second part, assume that (x_0,y_0) is a solution to (3.2). Then we parameterize its solutions as follows. First, note that

$$ax+my=b=ax_0+my_0.$$

Thus, $a(x-x_0)=m(y_0-y)$, or if we put $d=(a,m)$,

$$\frac{a}{d}(x-x_0)=\frac{m}{d}(y_0-y).$$

Since $(\frac{a}{d},\frac{m}{d})=1$, we obtain that for $k\in\mathbb{Z}$,

$$\begin{cases} x-x_0=k\cdot\frac{m}{d}, \\ y_0-y=k\cdot\frac{a}{d}, \end{cases} \qquad \Rightarrow \qquad \begin{cases} x=x_0+k\cdot\frac{m}{d}, \\ y=y_0-k\cdot\frac{a}{d}. \end{cases}$$

It turns out that x has only one possibility modulo $\frac{m}{d}$. ■

Now one may wonder if there is a way to construct an explicit expression of the solution to $ax\equiv b \pmod{m}$.

Definition 3.3 Let a and m be such that $(a,m)=1$. We say that $\bar{a}$ is a *modular inverse of a modulo m* if

$$a\bar{a}\equiv 1 \pmod{m}.$$

Theorem 3.9 Let a, b and m be such that $d\mid b$ where $d=(a,m)$. Then the solution to $ax\equiv b \pmod{m}$ is given by

$$x\equiv a'\cdot\frac{b}{d} \pmod{\frac{m}{d}},$$

where a' is the modular inverse of $\frac{a}{d}$ modulo $\frac{m}{d}$.

Proof. Note that we may rewrite $ax\equiv b \pmod{m}$ as

$$d\cdot\frac{a}{d}x\equiv d\cdot\frac{b}{d} \pmod{m},$$

which is equivalent to

$$\frac{a}{d}x\equiv\frac{b}{d} \pmod{\frac{m}{d}}$$

by Theorem 3.4 as $(d,m)=d$. Note also that $a'\cdot\frac{a}{d}\equiv 1 \pmod{\frac{m}{d}}$. Thus,

$$x\equiv a'\cdot\frac{b}{d} \pmod{\frac{m}{d}},$$

which is our desired result. ■

■ **Example 3.3** Consider $10x\equiv 15 \pmod{35}$: We have $d=(10,35)=5$. Meanwhile, $\frac{10}{5}\times 4\equiv 1 \pmod{\frac{35}{5}}$. Therefore, $x\equiv 4\times\frac{15}{5}=12 \pmod{\frac{35}{5}}$, i.e. $x\equiv 5 \pmod{7}$. ■

3.4 Chinese remainder theorem

We have seen that linear congruences are essentially equivalent to $x\equiv c \pmod{m}$.

Theorem 3.10 The system

$$x\equiv c_1 \pmod{m_1}, \tag{3.3a}$$
$$x\equiv c_2 \pmod{m_2}, \tag{3.3b}$$

has a solution if and only if $(m_1,m_2)\mid(c_2-c_1)$. The solution, if it exists, is unique modulo $[m_1,m_2]$.

Proof. From (3.3a), we may write $x = m_1 y + c_1$ for some indeterminate y. Substituting it into (3.3b), we have

$$m_1 y + c_1 \equiv c_2 \pmod{m_2},$$

or

$$m_1 y \equiv c_2 - c_1 \pmod{m_2}.$$

By Theorem 3.8, it is solvable if and only if $(m_1, m_2) \mid (c_2 - c_1)$. Further, the solution y is unique modulo $\frac{m_2}{(m_1,m_2)}$, and thus the solution x is unique modulo $m_1 \cdot \frac{m_2}{(m_1,m_2)} = [m_1, m_2]$ by Theorem 2.12. ■

Corollary 3.11 Let m_1 and m_2 be such that $(m_1, m_2) = 1$. Then the system in Theorem 3.10 is solvable, and its solution is unique modulo $m_1 m_2$.

In general, we may consider an analogous system with multiple linear congruences. Along this line, we have the *Chinese Remainder Theorem*, which first appeared in the writings of Sun Tzu (孙武: 孙子兵法), an ancient Chinese philosopher who lived during the Eastern Zhou period, and was further developed by the Chinese mathematician Qin Jiushao (秦九韶).

Theorem 3.12 (Chinese Remainder Theorem). Let $m_1, \ldots, m_r$ be such that $(m_i, m_j) = 1$ for $i \neq j$. Then the system $x \equiv c_i \pmod{m_i}$ for $1 \le i \le r$ has a unique solution modulo $m_1 \cdots m_r$.

Proof. This result follows from an iterative application of Corollary 3.11. ■

4. Fermat–Euler Theorem

4.1 Reduced residue systems

Definition 4.1 A set $\{a_1, a_2, \dots, a_h\}$ is called a *reduced residue system modulo m*, or a *reduced system modulo m*, if

(i) $a_i \not\equiv a_j \pmod{m}$ for any $i \neq j$;

(ii) $(a_i, m) = 1$ for $1 \le i \le h$;

(iii) For any integer a with $(a, m) = 1$, there exists an index i such that $a \equiv a_i \pmod{m}$.

Example 4.1 **(i).** $\{1, 5\}$ is a reduced system modulo 6; **(ii).** $\{1, 2, \dots, p-1\}$ is a reduced system modulo p for p a prime.

Theorem 4.1 Let $\{a_1, \dots, a_h\}$ be a reduced system modulo m and let k be an integer with $(k, m) = 1$. Then $\{ka_1, \dots, ka_h\}$ is also a reduced system modulo m.

Proof. Our proof is similar to that for Theorem 3.6.

(i). The same as Part (i) in the proof of Theorem 3.6.

(ii). Show $(ka_i, m) = 1$ for $1 \le i \le h$. Since k and a_i have no common divisors greater than 1 with m, so does their product ka_i.

(iii). Show $a \equiv ka_i \pmod{m}$ for a certain index i for any a with $(a, m) = 1$. Since $(k, m) = 1$, we may find an integer k' with $kk' \equiv 1 \pmod{m}$. Note that $(k', m) = 1$ for if d is a common divisor of k' and m, then $d \mid (kk' - mx) = 1$ where x is such that $kk' - 1 = mx$. Thus, $(ak', m) = 1$. Choose the index i such that $a_i \equiv ak' \pmod{m}$. Then $ka_i \equiv k(ak') = a(kk') \equiv a \pmod{m}$. ■

4.2 Euler's totient function

Note that a reduced system modulo m is a subset of a complete system modulo m. In particular, the size h of any reduced system modulo m equals the number of integers among $\{1, 2, \dots, m\}$ that are coprime to m.

Definition 4.2 Let n be a positive integer. *Euler's totient function* $\phi(n)$ denotes the number of integers among $\{1, 2, \dots, n\}$ that are coprime to n.

R The totient function was introduced by the Swiss mathematician Leonhard Euler

(*Novi commentarii academiae scientiarum imperialis Petropolitanae* **8** (1763), 74–104).

■ **Example 4.2** **(i).** $\phi(1) = 1$ for 1 is the only integer in $\{1\}$ that is coprime to 1; **(ii).** $\phi(3) = 2$ for 1 and 2 are the integers in $\{1,2,3\}$ that are coprime to 3; **(iii).** $\phi(6) = 2$ for 1 and 5 are the integers in $\{1,2,3,4,5,6\}$ that are coprime to 6. ■

R We may replace $\{1,2,\dots,n\}$ in the definition of Euler's totient function by any complete system modulo n.

Theorem 4.2 Let p be a prime and k be a positive integer. Then

$$\phi(p^k) = p^k - p^{k-1}. \tag{4.1}$$

Proof. Recall that $\phi(p^k)$ equals the number of integers in $\{1,\dots,p^k\}$ that are coprime to p^k, or in other words, that are not divisible by p. Since there are exactly p^{k-1} integers among $\{1,\dots,p^k\}$ that are multiples of p, namely, $p \cdot 1$, $p \cdot 2$, ..., $p \cdot p^{k-1}$, we have $\phi(p^k) = p^k - p^{k-1}$. ■

How to determine $\phi(n)$ if n is not a prime power?

Theorem 4.3 Let m and n be such that $(m,n) = 1$. Then

$$\phi(mn) = \phi(m)\phi(n). \tag{4.2}$$

Proof. We have shown in Theorem 3.7 that $\{bm + an : 1 \le a \le m, 1 \le b \le n\}$ is a complete system modulo mn. Thus, to compute $\phi(mn)$, it suffices to count the number of $bm + an$ such that $(bm + an, mn) = 1$. Note that

$$\begin{aligned}(bm+an,mn)=1 \quad &\Leftrightarrow \quad (bm+an,m)=1 \;\&\; (bm+an,n)=1\\ &\Leftrightarrow \quad (an,m)=1 \qquad\quad \&\; (bm,n)=1\\ &\Leftrightarrow \quad (a,m)=1 \qquad\quad\; \&\; (b,n)=1.\end{aligned}$$

Hence, there are $\phi(m)$ possibilities of a and $\phi(n)$ possibilities of b, and therefore $\phi(m)\phi(n)$ possibilities of admissible $bm + an$. It follows that $\phi(mn) = \phi(m)\phi(n)$. ■

R Given a function $f : \mathbb{Z}_{>0} \to \mathbb{C}$, we say that it is *multiplicative* if $f(1) = 1$ and for any m and n with $(m,n) = 1$,

$$f(mn) = f(m)f(n).$$

Corollary 4.4 For $n \ge 2$,

$$\phi(n) = n \cdot \prod_{p|n} \left(1 - \frac{1}{p}\right), \tag{4.3}$$

where the product runs over all prime divisors of n.

Proof. We write n in its canonical form $n = \prod_{i=1}^r p_i^{\alpha_i}$. Then by Theorem 4.3,

$$\phi(n) = \prod_{i=1}^r \phi(p_i^{\alpha_i}).$$

Further, making use of Theorem 4.2 gives

$$\prod_{i=1}^{r}\phi(p_i^{\alpha_i})=\prod_{i=1}^{r}(p_i^{\alpha_i}-p_i^{\alpha_i-1})=\prod_{i=1}^{r}p_i^{\alpha_i}\left(1-\frac{1}{p_i}\right)=\prod_{i=1}^{r}p_i^{\alpha_i}\cdot\prod_{i=1}^{r}\left(1-\frac{1}{p_i}\right)=n\cdot\prod_{i=1}^{r}\left(1-\frac{1}{p_i}\right),$$

thereby implying the desired result. ■

Theorem 4.5 For $n\geq 1$,

$$\sum_{d\mid n}\phi(d)=n, \tag{4.4}$$

where the sum runs over all positive divisors of n.

Proof. The formula is trivial when $n=1$. For $n>1$, we write n in the canonical form $n=p_1^{\alpha_1}\cdots p_r^{\alpha_r}$. Then all divisors of n are of the form $p_1^{\beta_1}\cdots p_r^{\beta_r}$ with $0\leq\beta_k\leq\alpha_k$ for each k. Thus,

$$\begin{aligned}\sum_{d\mid n}\phi(d)&=\sum_{\beta_1=0}^{\alpha_1}\cdots\sum_{\beta_r=0}^{\alpha_r}\phi(p_1^{\beta_1}\cdots p_r^{\beta_r})=\sum_{\beta_1=0}^{\alpha_1}\cdots\sum_{\beta_r=0}^{\alpha_r}\phi(p_1^{\beta_1})\cdots\phi(p_r^{\beta_r})\\&=\prod_{k=1}^{r}\left(\phi(1)+\phi(p_k)+\cdots+\phi(p_k^{\alpha_k})\right)\\&=\prod_{k=1}^{r}\left(1+(p_k-1)+(p_k^2-p_k)+\cdots+(p_k^{\alpha_k}-p_k^{\alpha_k-1})\right)\\&=\prod_{k=1}^{r}p_k^{\alpha_k}=n,\end{aligned}$$

as required. ■

R This relation can also be understood as follows. Consider the n fractions $\frac{1}{n}$, $\frac{2}{n}$, …, $\frac{n}{n}$. For each $\frac{k}{n}$, we may uniquely write it in the irreducible expression $\frac{k}{n}=\frac{a}{d}$ with $(a,d)=1$. Note that $d\mid n$. Also, since $1\leq k\leq n$, we have $1\leq a\leq d$. As there are exactly $\phi(d)$ such $\frac{a}{d}$, and they correspond to exactly $\phi(d)$ fractions among $\{\frac{k}{n}:1\leq k\leq n\}$, it follows that $n=\sum_{d\mid n}\phi(d)$.

4.3 Fermat–Euler Theorem

Theorem 4.6 (Fermat–Euler Theorem). If $(a,m)=1$, then

$$a^{\phi(m)}\equiv 1 \pmod{m}. \tag{4.5}$$

Proof. Let $\{x_1,\ldots,x_{\phi(m)}\}$ be a reduced system modulo m. Thus, $(x_i,m)=1$ for each i. Since $(a,m)=1$, we know from Theorem 4.1 that $\{ax_1,\ldots,ax_{\phi(m)}\}$ is also a reduced system modulo m. Thus,

$$\prod_{i=1}^{\phi(m)}x_i\equiv\prod_{i=1}^{\phi(m)}(ax_i)=a^{\phi(m)}\prod_{i=1}^{\phi(m)}x_i \pmod{m}.$$

Since $(x_i,m)=1$ for each i, we have $(\prod_i x_i,m)=1$. Therefore, by Corollary 3.5, $a^{\phi(m)}\equiv 1 \pmod{m}$. ■

The case where $m=p$ is a prime is also known as *Fermat's Theorem.*

Corollary 4.7 (Fermat's Theorem). If p is a prime and $p \nmid a$, then

$$a^{p-1} \equiv 1 \pmod{p}. \tag{4.6}$$

4.4 Binomial coefficients

Definition 4.3 For integers $m \geq n \geq 0$, the *binomial coefficients* are defined by

$$\binom{m}{n} = \frac{m!}{n!(m-n)!} = \frac{m(m-1)\cdots(m-n+1)}{n(n-1)\cdots 1}.$$

In particular, $\binom{m}{0} = 1$.

Theorem 4.8 (Pascal's Identity). For integers $m \geq n > 0$,

$$\binom{m+1}{n} = \binom{m}{n} + \binom{m}{n-1}. \tag{4.7}$$

Proof. We have

$$\begin{aligned}
\binom{m}{n} + \binom{m}{n-1} &= \frac{m!}{n!(m-n)!} + \frac{m!}{(n-1)!(m-n+1)!} \\
&= \frac{m!}{(n-1)!(m-n)!} \cdot \frac{1}{n} + \frac{m!}{(n-1)!(m-n)!} \cdot \frac{1}{m-n+1} \\
&= \frac{m!}{(n-1)!(m-n)!} \cdot \frac{m+1}{n(m-n+1)} \\
&= \frac{(m+1)!}{(n)!(m-n+1)!},
\end{aligned}$$

which is exactly $\binom{m+1}{n}$. ■

Theorem 4.9 (Binomial Theorem). For $n \geq 0$,

$$(x+y)^n = \sum_{r=0}^{n} \binom{n}{r} x^r y^{n-r}. \tag{4.8}$$

Proof. We argue by induction on n. When $n = 0$, both sides of (4.8) are 1, and when $n = 1$, both sides of (4.8) are $x + y$. Assuming that (4.8) is true for some $n \geq 1$, we want to show that it is also true for $n + 1$. Note that

$$\begin{aligned}
(x+y)^{n+1} &= (x+y)(x+y)^n \\
&= (x+y)\left(\sum_{r=0}^{n} \binom{n}{r} x^r y^{n-r}\right) \\
&= \sum_{r=0}^{n} \binom{n}{r} x^{r+1} y^{n-r} + \sum_{r=0}^{n} \binom{n}{r} x^r y^{n-r+1} \\
&= \left(x^{n+1} + \sum_{r=0}^{n-1} \binom{n}{r} x^{r+1} y^{n-r}\right) + \left(y^{n+1} + \sum_{r=1}^{n} \binom{n}{r} x^r y^{n-r+1}\right) \\
&= \left(x^{n+1} + \sum_{r=1}^{n} \binom{n}{r-1} x^r y^{n-r+1}\right) + \left(y^{n+1} + \sum_{r=1}^{n} \binom{n}{r} x^r y^{n-r+1}\right)
\end{aligned}$$

$$
\begin{aligned}
&= x^{n+1} + y^{n+1} + \sum_{r=1}^{n} \left(\binom{n}{r-1} + \binom{n}{r} \right) x^r y^{n-r+1} \\
&= x^{n+1} + y^{n+1} + \sum_{r=1}^{n} \binom{n+1}{r} x^r y^{n-r+1} \\
&= \sum_{r=0}^{n+1} \binom{n+1}{r} x^r y^{n-r+1},
\end{aligned}
$$

which is exactly the $n+1$ case of (4.8). ■

Corollary 4.10 The binomial coefficients $\binom{m}{n}$ are integers.

Theorem 4.11 Let p be a prime. Given any nonzero integer n, we denote by $\nu_p(n)$ the unique nonnegative integer k such that $p^k \mid n$ and $p^{k+1} \nmid n$, namely, $\nu_p(n)$ is the power of p in the canonical form of n. Let α be a positive integer. For $1 \le r \le p^\alpha$,

$$\nu_p\left(\binom{p^\alpha}{r} \right) = \alpha - \nu_p(r). \tag{4.9}$$

In particular, for any r with $1 \le r \le p-1$, we have $p \mid \binom{p}{r}$.

Proof. Recall that $\binom{p^\alpha}{r} = \frac{p^\alpha(p^\alpha-1)\cdots(p^\alpha-r+1)}{r(r-1)\cdots 1}$. For each s with $1 \le s \le r-1 < p^\alpha$, we observe the simple fact that $\nu_p(s) = \nu_p(p^\alpha - s)$. Hence, $\nu_p(\binom{p^\alpha}{r}) = \nu_p(p^\alpha) - \nu_p(r) = \alpha - \nu_p(r)$. ■

Theorem 4.11 has two important consequences.

Theorem 4.12 For $\alpha \ge 1$ and p prime, if

$$m \equiv 1 \pmod{p^\alpha},$$

then

$$m^p \equiv 1 \pmod{p^{\alpha+1}}.$$

Proof. We write $m = kp^\alpha + 1$ for a certain integer k. Then

$$m^p = (kp^\alpha + 1)^p = \sum_{r=0}^{p} \binom{p}{r} (kp^\alpha)^r = 1 + \sum_{r=1}^{p} \binom{p}{r} (kp^\alpha)^r.$$

Now, for $1 \le r \le p$, $\binom{p}{r} \cdot (p^\alpha)^r$ is always divisible by $p^{\alpha+1}$. ■

Theorem 4.13 For $k \ge 1$ and p prime,

$$(x_1 + x_2 + \cdots + x_k)^p \equiv x_1^p + x_2^p + \cdots + x_k^p \pmod{p}. \tag{4.10}$$

Proof. We apply induction on k. The $k=1$ case is trivial. Assume that the statement is true for some $k \ge 1$. Then we prove the $k+1$ case:

$$
\begin{aligned}
(x_1 + x_2 + \cdots + x_{k+1})^p &= \big(x_1 + (x_2 + \cdots + x_{k+1})\big)^p \\
&= \sum_{r=0}^{p} \binom{p}{r} x_1^r (x_2 + \cdots + x_{k+1})^{p-r} \\
&\equiv x_1^p + (x_2 + \cdots + x_{k+1})^p
\end{aligned}
$$

$$\equiv x_1^p + x_2^p + \cdots + x_{k+1}^p \pmod p,$$

by our inductive assumption. ■

4.5 Euler's proof of the Fermat–Euler Theorem

Let us close this lecture with Euler's proof of the Fermat–Euler Theorem. Note that the case where $m = 1$ is trivial. We start by showing that for $\alpha \geq 1$ and p prime, if a is such that $(a, p) = 1$,

$$a^{\phi(p^\alpha)} \equiv 1 \pmod{p^\alpha}. \tag{4.11}$$

For its proof, we first choose $k = a$ in Theorem 4.13 and then put $x_1 = \cdots = x_a = 1$. Thus, $a^p \equiv a \pmod p$. Since $(a, p) = 1$, we have $a^{p-1} \equiv 1 \pmod p$. Now, by an iterative application of Theorem 4.12, we have $a^{(p-1)p} \equiv 1 \pmod{p^2}$, ..., and $a^{(p-1)p^{\alpha-1}} \equiv 1 \pmod{p^\alpha}$, which is exactly (4.11).

Now, for integers $m \geq 2$, we write $m = \prod_i p_i^{\alpha_i}$. Assume that a is such that $(a, m) = 1$, and hence that $(a, p_i) = 1$ for each i. We also write for convenience $m = p_i^{\alpha_i} m_i$. Since ϕ is multiplicative, $\phi(m) = \phi(p_i^{\alpha_i})\phi(m_i)$. Thus, by (4.11),

$$a^{\phi(m)} = \left(a^{\phi(p_i^{\alpha_i})}\right)^{\phi(m_i)} \equiv 1^{\phi(m_i)} = 1 \pmod{p_i^{\alpha_i}}.$$

Since these $p_i^{\alpha_i}$ are pairwise coprime while $m = \prod_i p_i^{\alpha_i}$, we know from the Chinese Remainder Theorem that

$$a^{\phi(m)} \equiv 1 \pmod m,$$

as desired.

5. Primitive roots

5.1 Powers of integers

Let m be a positive integer and a be an integer with $(a,m)=1$. Let $k \geq 0$ be a nonnegative integer.

(i) For nonnegative powers of a, we know that a^k is an integer, and hence we may directly determine the residue class of a^k modulo m.

(ii) For negative powers of a, we recall from Definition 3.3 that there exists an integer $\overline{a}$ such that $a\overline{a} \equiv 1 \pmod{m}$. Thus, we may use a^{-1} to represent the residue class of $\overline{a}$ modulo m. In particular, we have $aa^{-1} \equiv 1 \pmod{m}$, which is a natural analogy to the usual inverse of integers; this explains why we call a^{-1} the modular inverse of a in Definition 3.3. Now we may naturally define negative powers of a modulo m by $a^{-k} \equiv (a^{-1})^k \pmod{m}$.

R Note that if a is such that $(a,m)>1$, then there is no integer $\overline{a}$ such that $a\overline{a} \equiv 1 \pmod{m}$, since by Theorem 2.5, $ax-1=my$ has no integer solutions x and y. Thus, we cannot define the negative powers of a modulo m in this case. However, nonnegative powers of a can still be defined as normal powers.

From the above definition, we have the following trivial fact.

Theorem 5.1 Let m be a positive integer and a,b be integers with $(a,m)=(b,m)=1$ and $a \equiv b \pmod{m}$. Then for any integer x,

$$a^x \equiv b^x \pmod{m}. \tag{5.1}$$

The next two results show that integer powers in the modular sense have similar properties to normal powers of integers.

Theorem 5.2 Let m be a positive integer and a,b be integers with $(a,m)=(b,m)=1$. Then for any integer x,

$$(ab)^x \equiv a^x b^x \pmod{m}. \tag{5.2}$$

Proof. If $x \geq 0$, then $(ab)^x = a^x b^x$ as normal integer powers, and hence they are congruent modulo m. If $x<0$, we first note that $(ab)^{-1}\equiv a^{-1}b^{-1}\pmod{m}$ for

$$(ab)\cdot(a^{-1}b^{-1})=(aa^{-1})\cdot(bb^{-1})\equiv 1\cdot 1=1\pmod{m}.$$

Thus,

$$(ab)^x\equiv\left((ab)^{-1}\right)^{-x}\equiv(a^{-1}b^{-1})^{-x}=(a^{-1})^{-x}(b^{-1})^{-x}\equiv a^xb^x\pmod{m},$$

as desired. ■

Theorem 5.3 Let m be a positive integer and a be an integer with $(a,m)=1$. Then

(i) $1^{-1}\equiv 1\pmod{m}$;

(ii) $(a^{-1})^{-1}\equiv a\pmod{m}$;

(iii) For any integers x and y, we have $a^{x+y}\equiv a^xa^y\pmod{m}$;

(iv) For any integers x and y, we have $a^{xy}\equiv(a^x)^y\pmod{m}$.

Proof. **(i).** Note that $1\cdot 1\equiv 1\pmod{m}$, and hence that $1^{-1}\equiv 1\pmod{m}$.

(ii). Note that a^{-1} is the modular inverse of a modulo m and vice versa by definition. This means that $(a^{-1})^{-1}\equiv a\pmod{m}$.

(iii). This relation is trivial if x and y are simultaneously nonnegative, or simultaneously nonpositive. Without loss of generality, we assume that $x>0>y$. In particular, we may further assume that $x+y\geq 0$, for if $x+y<0$, we only need to rewrite the congruence as $(a^{-1})^{-(x+y)}\equiv(a^{-1})^{-x}(a^{-1})^{-y}\pmod{m}$. Now, we note that $a^x=a^{x+y-y}=a^{x+y}a^{-y}$ for both $x+y$ and $-y$ are nonnegative integers. Hence,

$$a^x\cdot a^y=(a^{x+y}a^{-y})\cdot a^y\equiv(a^{x+y}a^{-y})\cdot(a^{-1})^{-y}=a^{x+y}\cdot(a\cdot a^{-1})^{-y}\equiv a^{x+y}\cdot 1^{-y}=a^{x+y}\pmod{m}.$$

(iv). We require three basic facts. Firstly, for x and y nonnegative integers,

$$(a^x)^y=a^{xy};\tag{5.3}$$

this is a property of normal integer powers. Secondly, for x a nonnegative integer,

$$(a^{-1})^x\equiv a^{-x}\pmod{m};\tag{5.4}$$

this follows from the definition of negative powers in the modular sense. Thirdly, for x an integer,

$$(a^x)^{-1}\equiv a^{-x}\pmod{m};\tag{5.5}$$

this follows from Part (iii) as $a^xa^{-x}\equiv a^{x+(-x)}=a^0=1\pmod{m}$, namely, a^{-x} is the modular inverse of a^x. Now, we prove Part (iv) according to the following four cases. **(a).** If $x,y\geq 0$, then by (5.3) $a^{xy}=(a^x)^y$ and thus they are congruent modulo m. **(b).** If $x\geq 0>y$, then

$$(a^x)^y\overset{(5.4)}{\equiv}\left((a^x)^{-1}\right)^{-y}\overset{(5.5)}{\equiv}(a^{-x})^{-y}\overset{(5.4)}{\equiv}\left((a^{-1})^x\right)^{-y}\overset{(5.3)}{=}(a^{-1})^{-xy}\overset{(5.4)}{\equiv}a^{xy}\pmod{m}.$$

(c). If $y\geq 0>x$, then

$$(a^x)^y\overset{(5.4)}{\equiv}\left((a^{-1})^{-x}\right)^y\overset{(5.3)}{=}(a^{-1})^{-xy}\overset{(5.4)}{\equiv}a^{xy}\pmod{m}.$$

(d). If $x,y<0$, then

$$(a^x)^y\overset{(5.4)}{\equiv}\left((a^x)^{-1}\right)^{-y}\overset{(5.5)}{\equiv}(a^{-x})^{-y}\overset{(5.3)}{=}a^{xy}\pmod{m}.$$

The desired result hence holds. ■

5.2 Orders

By the Fermat–Euler Theorem (Theorem 4.6), we have $a^{\phi(m)} \equiv 1 \pmod{m}$, indicating that there exists at least one positive integer x such that $a^x \equiv 1 \pmod{m}$.

Definition 5.1 Let m be a positive integer and a be an integer with $(a,m) = 1$. The smallest positive integer d such that

$$a^d \equiv 1 \pmod{m} \tag{5.6}$$

is called the *order of a modulo m*, denoted by $\operatorname{ord}_m a$.

■ **Example 5.1** **(i).** We have $\operatorname{ord}_5 2 = 4$ for $2^1 \equiv 2$, $2^2 \equiv 4$, $2^3 \equiv 3$ and $2^4 \equiv 1 \pmod{5}$. **(ii).** We have $\operatorname{ord}_7 2 = 3$ for $2^1 \equiv 2$, $2^2 \equiv 4$ and $2^3 \equiv 1 \pmod{7}$. ■

R By definition, it is immediate that if a and b are such that $(a,m) = (b,m) = 1$ and that $a \equiv b \pmod{m}$, then $\operatorname{ord}_m a = \operatorname{ord}_m b$.

Theorem 5.4 Let m be a positive integer and a be an integer with $(a,m) = 1$. Then an integer x satisfies $a^x \equiv 1 \pmod{m}$ if and only if $\operatorname{ord}_m a \mid x$. In particular, $\operatorname{ord}_m a \mid \phi(m)$.

Proof. Let $d = \operatorname{ord}_m a$. Then $a^d \equiv 1 \pmod{m}$ by definition. If $d \mid x$, then we may write $x = q \cdot d$ and thus,

$$a^x = a^{qd} \equiv (a^d)^q \equiv 1^q \equiv 1 \pmod{m}.$$

Assume that there exists an x with $d \nmid x$ such that $a^x \equiv 1 \pmod{m}$. Thus, we may write $x = q \cdot d + r$ for q and r integers with $0 < r < d$. It follows that

$$1 \equiv a^x = a^{qd+r} \equiv a^{qd} \cdot a^r \equiv (a^d)^q \cdot a^r \equiv 1 \cdot a^r = a^r \pmod{m}.$$

But this violates the assumption that d is the smallest positive integer such that $a^d \equiv 1 \pmod{m}$. Finally, $\operatorname{ord}_m a \mid \phi(m)$ since $a^{\phi(m)} \equiv 1 \pmod{m}$ by the Fermat–Euler Theorem. ■

Theorem 5.5 Let m be a positive integer and a be an integer with $(a,m) = 1$. If we write $d = \operatorname{ord}_m a$, then for any integer k,

$$\operatorname{ord}_m a^k = \frac{d}{(d,k)}. \tag{5.7}$$

In particular, for any positive d^* with $d^* \mid d$, we have $\operatorname{ord}_m a^{\frac{d}{d^*}} = d^*$.

Proof. We write $d' = \operatorname{ord}_m a^k$ and $\delta = (d,k)$. First, noting that $(a^k)^{\frac{d}{\delta}} \equiv (a^d)^{\frac{k}{\delta}} \equiv 1^{\frac{k}{\delta}} \equiv 1 \pmod{m}$, we have $d' \mid \frac{d}{\delta}$ by Theorem 5.4. Also, $a^{kd'} \equiv (a^k)^{d'} \equiv 1 \pmod{m}$, and therefore $d \mid kd'$ by Theorem 5.4, thereby implying that $\frac{d}{\delta} \mid \frac{k}{\delta} d'$. Further, we have $(\frac{d}{\delta}, \frac{k}{\delta}) = 1$ since $\delta = (d,k)$. Hence, $\frac{d}{\delta} \mid d'$. It follows that $d' = \frac{d}{\delta}$. Finally, we choose $k = \frac{d}{d^*}$ and note that $(d, \frac{d}{d^*}) = \frac{d}{d^*}$, thereby getting the last part. ■

Theorem 5.6 Let m be a positive integer and a,b be integers with $(a,m) = (b,m) = 1$. Let $d_a = \operatorname{ord}_m a$ and $d_b = \operatorname{ord}_m b$. If $(d_a, d_b) = 1$, then $\operatorname{ord}_m (ab) = d_a d_b$.

Proof. Let $d = \operatorname{ord}_m(ab)$. First, noting that $(ab)^{d_a d_b} \equiv (a^{d_a})^{d_b} \cdot (b^{d_b})^{d_a} \equiv 1^{d_b} \cdot 1^{d_a} \equiv 1 \pmod{m}$, we have $d \mid d_a d_b$. Also, $a^{dd_b} \equiv a^{dd_b} \cdot 1^d \equiv a^{dd_b} \cdot (b^{d_b})^d \equiv (ab)^{dd_b} \equiv \big((ab)^d\big)^{d_b} \equiv 1^{d_b} \equiv 1 \pmod{m}$, and thus $d_a \mid dd_b$. Noting further that $(d_a, d_b) = 1$, we have $d_a \mid d$. Similarly, $d_b \mid d$ and thus $d_a d_b \mid d$ since $(d_a, d_b) = 1$. It follows that $d = d_a d_b$. ■

Theorem 5.7 Let m be a positive integer and a, b be integers with $(a,m) = (b,m) = 1$. Let $d_a = \operatorname{ord}_m a$ and $d_b = \operatorname{ord}_m b$. There exists an integer c with $(c,m) = 1$ such that $\operatorname{ord}_m c = \operatorname{lcm}(d_a, d_b)$.

Proof. We write in the canonical form $d_a = \prod_i p_i^{\alpha_i}$ and $d_b = \prod_i p_i^{\beta_i}$ with $\alpha_i, \beta_i \geq 0$. Define

$$d_1 = \prod_{k:\alpha_k > \beta_k} p_k^{\alpha_k} \qquad \text{and} \qquad d_2 = \prod_{\ell:\beta_\ell \geq \alpha_\ell} p_\ell^{\beta_\ell}.$$

Then $d_1 \mid d_a$, $d_2 \mid d_b$, $(d_1, d_2) = 1$ and $d_1 d_2 = \operatorname{lcm}(d_a, d_b)$. By Theorem 5.5, we have

$$\operatorname{ord}_m a^{\frac{d_a}{d_1}} = d_1 \qquad \text{and} \qquad \operatorname{ord}_m b^{\frac{d_b}{d_2}} = d_2.$$

Now if we choose

$$c = a^{\frac{d_a}{d_1}} b^{\frac{d_b}{d_2}},$$

then Theorem 5.6 tells us that $\operatorname{ord}_m c = d_1 d_2 = \operatorname{lcm}(d_a, d_b)$. ■

Theorem 5.8 Let m be a positive integer and $\{a_1, a_2, \ldots, a_{\phi(m)}\}$ be a reduced residue system modulo m. Let $d_i = \operatorname{ord}_m a_i$ for $1 \leq i \leq \phi(m)$ and define $D = \max_{1 \leq i \leq \phi(m)}\{d_i\}$. Then $D \mid \phi(m)$, and $d_i \mid D$ for each $1 \leq i \leq \phi(m)$.

Proof. First, $D \mid \phi(m)$ follows from Theorem 5.4 and the fact that D is the order of a certain a_i. For the second part, we argue by contradiction. Assume that there exists a certain a_j of order $d = \operatorname{ord}_m a_j$ such that $d \nmid D$. Note that for this d, we have $\operatorname{lcm}(d, D) > D$. Then by Theorem 5.7, we get an integer of order $\operatorname{lcm}(d, D) > D$. But this violates the fact that D is the maximum among the orders. ■

5.3 Primitive roots

Recall that the orders modulo m are always divisors of $\phi(m)$. We now focus on the case where the order equals $\phi(m)$.

Definition 5.2 An integer g is called a *primitive root* of m if $\operatorname{ord}_m g = \phi(m)$.

Theorem 5.9 If m has a primitive root g, then $\{g, g^2, \ldots, g^{\phi(m)}\}$ gives a reduced residue system modulo m.

R If m has a primitive root, then the multiplicative group $(\mathbb{Z}/m\mathbb{Z})^\times$ is cyclic.

Proof. Note that the $\phi(m)$ integers g, ..., $g^{\phi(m)}$ are coprime to m since $(g,m) = 1$. Hence, it suffices to show that they are pairwise incongruent modulo m. Assume not; then there are integers i and j with $1 \leq i < j \leq \phi(m)$ such that $g^i \equiv g^j \pmod{m}$, or $g^{j-i} \equiv 1 \pmod{m}$. But g is a primitive root of m, and thus $\operatorname{ord}_m g = \phi(m)$. By Theorem 5.4, $\phi(m) \mid (j-i)$, which is absurd. ■

Theorem 5.10 If m has a primitive root, then there are $\phi(\phi(m))$ primitive roots among $1,2,\ldots,m$.

Proof. Let g be a primitive root of m and hence $\operatorname{ord}_m g = \phi(m)$. Then Theorem 5.9 tells us that the reduced system modulo m can be represented by $\{g,\ldots,g^{\phi(m)}\}$. Thus, it suffices to determine the number of i with $1 \le i \le \phi(m)$ such that $\operatorname{ord}_m g^i = \phi(m)$. On the other hand, we know from Theorem 5.5 that $\operatorname{ord}_m g^i = \frac{\phi(m)}{(i,\phi(m))}$. So we only need to count the number of i such that $(i,\phi(m)) = 1$ and there are $\phi(\phi(m))$ such i among $1,\ldots,\phi(m)$. ■

5.4 Lagrange's polynomial congruence theorem

Here, we present a theorem of the Italian mathematician Joseph-Louis Lagrange, which will be a key for confirming the existence of primitive roots of an odd prime.

Theorem 5.11 (Lagrange's Polynomial Congruence Theorem). Let p be a prime. Let $f(x) = a_n x^n + \cdots + a_1 x + a_0$ be a polynomial with integer coefficients such that $p \nmid a_n$. Then the congruence

$$f(x) \equiv 0 \pmod{p}$$

has at most n solutions modulo p.

Proof. We argue by induction on the degree n of $f(x)$. When $n = 1$, $f(x)$ is linear and the statement is trivial. Now we assume that the statement is true for $1,\ldots,n$ with $n \ge 1$. Let $f(x)$ be of degree $n+1$. If $f(x) \equiv 0 \pmod{p}$ has no solutions, then there is nothing to prove. If there is one solution, say $x \equiv x_0 \pmod{p}$, then $f(x_0) \equiv 0 \pmod{p}$. Now, we consider $g(x) = f(x) - f(x_0) = (x - x_0)q(x)$ where $q(x)$ is a polynomial with integer coefficients whose degree is n. Note that $f(x) \equiv 0 \pmod{p}$ is equivalent to $g(x) \equiv 0 \pmod{p}$. Since p is a prime, we either have $x - x_0 \equiv 0 \pmod{p}$ which has one solution modulo p, or $q(x) \equiv 0 \pmod{p}$ which has at most n solutions modulo p by our inductive assumption. It follows that there are at most $n+1$ solutions to $f(x) \equiv 0 \pmod{p}$, as desired. ■

5.5 Existence of primitive roots

Now we are in a position to characterize which integers have primitive roots.

Theorem 5.12 Every odd prime p has a primitive root.

Proof. As in Theorem 5.8, we write $d_k = \operatorname{ord}_p k$ for $1 \le k \le p-1$, and define $D = \max_k\{d_k\}$ so that $D \mid \phi(p) = p-1$. Since $d_k \mid D$, we have $k^D \equiv 1 \pmod{p}$ for each k. It turns out that the congruence $x^D - 1 \equiv 0 \pmod{p}$ has $p-1$ solutions modulo p. By Lagrange's Polynomial Congruence Theorem (Theorem 5.11), we have $D \ge p-1$. Combining with the fact that $D \mid p-1$, we have $D = p-1$. Therefore, there exists an integer g of order $D = p-1 = \phi(p)$, thereby giving our desired primitive root. ■

Lemma 5.13 For any odd prime p, there exists a primitive root g such that $p \mid (g^{p-1} - 1)$ and $p^2 \nmid (g^{p-1} - 1)$.

Proof. Let g be an arbitrary primitive root of p. Then $g^{p-1} \equiv 1 \pmod{p}$, namely, $p \mid (g^{p-1} - 1)$. If we also have $p^2 \nmid (g^{p-1} - 1)$, there is nothing to prove. If $p^2 \mid (g^{p-1} - 1)$, namely, $g^{p-1} - 1 \equiv 0 \pmod{p^2}$, then we note that $g_* = p + g$ is also a primitive root of p.

Meanwhile,

$$\begin{aligned} g_*^{p-1} - 1 = (p+g)^{p-1} - 1 &= \sum_{r=0}^{p-1} \binom{p-1}{r} p^r g^{p-1-r} - 1 \\ &\equiv g^{p-1} + p(p-1)g^{p-2} - 1 \equiv -pg^{p-2} \not\equiv 0 \pmod{p^2}. \end{aligned}$$

Hence, in this case g_* is the desired primitive root. ■

Theorem 5.14 For any odd prime p, let g be a primitive root as in Lemma 5.13. Then for any positive integer α, g is also a primitive root of p^α. In particular, p^α always has an odd primitive root.

Proof. Since g is a primitive root of p as in Lemma 5.13, we have $\operatorname{ord}_p g = \phi(p) = p-1$ and g is such that

$$g^{p-1} = px + 1$$

with $p \nmid x$. Let $\operatorname{ord}_{p^\alpha} g = d$. Then $g^d \equiv 1 \pmod{p^\alpha}$, and thus $g^d \equiv 1 \pmod{p}$. Hence, $(p-1) \mid d$. On the other hand, $d \mid \phi(p^\alpha) = (p-1)p^{\alpha-1}$. Hence, d is of the form $d = (p-1)p^s$ for some $0 \le s \le \alpha - 1$. Now, recalling that $p \nmid x$, we have, with an application of Theorem 4.11,

$$g^d = g^{(p-1)p^s} = (px+1)^{p^s} = \sum_{r=0}^{p^s} \binom{p^s}{r} (px)^r \equiv 1 + p^{s+1}x \not\equiv 1 \pmod{p^{s+2}}.$$

However, $g^d \equiv 1 \pmod{p^\alpha}$. Hence, $\alpha < s+2$. It follows that the only possibility is $s = \alpha - 1$, implying that $\operatorname{ord}_{p^\alpha} g = d = (p-1)p^{\alpha-1} = \phi(p^\alpha)$, namely, g is a primitive root of p^α. Finally, we observe that both g and $g + p^\alpha$ are primitive roots of p^α, and they are of different parities, thereby concluding the last part. ■

Theorem 5.15 For any odd prime p and positive integer α, let g be an odd primitive root of p^α. Then g is also a primitive root of $2p^\alpha$.

Proof. Note that g being an odd primitive root of p^α implies that $(g, 2p^\alpha) = 1$. Let $d = \operatorname{ord}_{2p^\alpha} g$ and we have $d \mid \phi(2p^\alpha)$. Then $g^d \equiv 1 \pmod{2p^\alpha}$, and hence, $g^d \equiv 1 \pmod{p^\alpha}$. Since g is a primitive root of p^α, we have $\phi(p^\alpha) = \operatorname{ord}_{p^\alpha} g \mid d$. However, $\phi(2p^\alpha) = \phi(p^\alpha) = (p-1)p^{\alpha-1}$. It follows that $d = \phi(2p^\alpha)$, namely, g is a primitive root of $2p^\alpha$. ■

Theorem 5.16 A positive integer m has a primitive root if and only if m is of the form 1, 2, 4, p^α or $2p^\alpha$ where p is an odd prime and α is a positive integer.

Proof. Note that 1 has a primitive root 1, that 2 has a primitive root 1, and that 4 has a primitive root 3. It remains to show that no other positive integers have primitive roots.

We first exclude integers m that can be written as $m = st$ with $s,t \ge 3$ and $(s,t) = 1$. Recall that Euler's totient function ϕ is multiplicative, namely, $\phi(m) = \phi(s)\phi(t)$. Also, $\phi(s)$ and $\phi(t)$ are even by using Theorem 4.2. Thus, $\frac{\phi(m)}{2}$ is a integer. We shall prove that for any a with $(a,m) = 1$, $a^{\frac{\phi(m)}{2}} \equiv 1 \pmod{m}$. To see this, we have

$$a^{\frac{\phi(m)}{2}} = \left(a^{\phi(s)}\right)^{\frac{\phi(t)}{2}} \equiv 1^{\frac{\phi(t)}{2}} = 1 \pmod{s},$$

and similarly,

$$a^{\frac{\phi(m)}{2}} \equiv 1 \pmod{t}.$$

Note that $(s,t) = 1$ and $st = m$. By the Chinese Remainder Theorem, we have $a^{\frac{\phi(m)}{2}} \equiv 1 \pmod{m}$. Hence, m has no primitive roots.

Finally, we exclude integers of the form 2^α with $\alpha \geq 3$. Note that if a is such that $(a, 2^\alpha) = 1$, then a is odd and we write $a = 2b+1$. We prove that $a^{\frac{\phi(2^\alpha)}{2}} = a^{2^{\alpha-2}} \equiv 1 \pmod{2^\alpha}$ always holds. To see this, we have, with Theorem 4.11 applied,

$$\begin{aligned} a^{\frac{\phi(2^\alpha)}{2}} &= (2b+1)^{2^{\alpha-2}} = \sum_{r=0}^{2^{\alpha-2}} \binom{2^{\alpha-2}}{r} (2b)^r \\ &\equiv 1 + 2^{\alpha-2}(2b) + (2^{\alpha-2}-1)2^{\alpha-3}(2b)^2 \\ &\equiv 1 + 2^{\alpha-1}(b-b^2) \\ &\equiv 1 \pmod{2^\alpha}. \end{aligned}$$

Hence 2^α has no primitive roots when $\alpha \geq 3$. ■

6. Quadratic residues

6.1 Quadratic residues

Assume that p is a prime and that x is such that $1 \le x \le p-1$. Recall from Theorem 4.1 that $\{xy : 1 \le y \le p-1\}$ forms a reduced system modulo p. Hence, for any integer a with $(a,p)=1$, there exists a unique x' with $1 \le x' \le p-1$ such that $xx' \equiv a \pmod{p}$.

Definition 6.1 We call x' the *associate of x with respect to a modulo p* if

$$xx' \equiv a \pmod{p}$$

with $1 \le x' \le p-1$.

We are in particular interested in the case where the associate of x is itself.

Definition 6.2 Let p be a prime and a be such that $(a,p)=1$. We say that a is a *quadratic residue modulo p* if there exists an x such that

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

If such x does not exist, we say that a is a *quadratic non-residue modulo p*.

R By definition, it is straightforward to see that for p a prime, if a and b are such that $(a,p)=(b,p)=1$ and $a \equiv b \pmod{p}$, then a and b are simultaneously quadratic residues modulo p, or simultaneously quadratic non-residues modulo p. In this case, we shall say that a and b are in the same quadratic residue or non-residue class modulo p.

Note that when $p=2$, for any a with $(a,2)=1$ so that a is an odd integer, we always have $a \equiv 1 = 1^2 \pmod{2}$. Thus, all odd integers are quadratic residues modulo 2. Below, we only focus on the case where $p \ge 3$.

Lemma 6.1 Let $p \ge 3$ be a prime and x_0 be such that $(x_0,p)=1$. Then

$$x^2 \equiv x_0^2 \pmod{p} \tag{6.1}$$

has exactly two solutions

$$x_+ \equiv x_0 \pmod{p} \qquad \text{and} \qquad x_- \equiv -x_0 \pmod{p},$$

and in particular $x_+ \not\equiv x_- \pmod{p}$.

Proof. We rewite (6.1) as

$$(x-x_0)(x+x_0) \equiv 0 \pmod{p}.$$

Since p is a prime, it follows that $p \mid (x-x_0)$ or $p \mid (x+x_0)$, thereby leading to the two solutions $x_\pm$. Also, $x_+ \not\equiv x_- \pmod{p}$; otherwise, we have $x_0 \equiv -x_0 \pmod{p}$, or $p \mid 2x_0$, or $p \mid x_0$ since $p \geq 3$ is a prime. But this violates the assumption that $(x_0, p) = 1$. ■

Theorem 6.2 Let $p \geq 3$ be a prime.

(i) If a is a quadratic residue modulo p, then there are exactly two distinct residue classes $x \equiv x_1, x_2$ modulo p with $x_2 \equiv -x_1 \pmod{p}$ such that $x^2 \equiv a \pmod{p}$.

(ii) There are exactly $\frac{p-1}{2}$ quadratic residue classes modulo p, and $\frac{p-1}{2}$ quadratic non-residue classes modulo p. In particular, the quadratic residue classes can be represented by $\{1^2, 2^2, \ldots, (\frac{p-1}{2})^2\}$ modulo p.

Proof. **(i).** Since a is a quadratic residue, we may always find an x_1 such that $x_1^2 \equiv a \pmod{p}$. Thus, by Lemma 6.1, the only two solutions to $x^2 \equiv a \equiv x_1^2 \pmod{p}$ are $x \equiv \pm x_1 \pmod{p}$ and they are distinct.

(ii). First, Part (i) implies that there are at most $\frac{p-1}{2}$ quadratic residue classes modulo p. Otherwise, if there are at least $\frac{p+1}{2}$ quadratic residue classes, then there are at least $2 \cdot \frac{p+1}{2} = p+1$ residue classes modulo p, which is impossible. Next, we show that $1^2, \ldots, (\frac{p-1}{2})^2$ are pairwise distinct residue classes modulo p. To see this, we choose $1 \leq i, j \leq \frac{p-1}{2}$ with $i \neq j$. We claim that $i^2 \not\equiv j^2 \pmod{p}$. Otherwise, if $i^2 \equiv j^2 \pmod{p}$, then $p \mid (i-j)(i+j)$. But since $1 \leq i, j \leq \frac{p-1}{2}$ and $i \neq j$, both $i-j$ and $i+j$ are not multiples of p, thereby leading to a contradiction. Thus, there are exactly $\frac{p-1}{2}$ quadratic residue classes modulo p, characterized by $\{1^2, \ldots, (\frac{p-1}{2})^2\}$ modulo p, and as a consequence, there are exactly $(p-1) - \frac{p-1}{2} = \frac{p-1}{2}$ quadratic non-residue classes modulo p. ■

Theorem 6.3 Let $p \geq 3$ be a prime.

(i) If a is a quadratic residue modulo p, then

$$(p-1)! \equiv -a^{\frac{p-1}{2}} \pmod{p}. \tag{6.2}$$

(ii) If a is a quadratic non-residue modulo p, then

$$(p-1)! \equiv a^{\frac{p-1}{2}} \pmod{p}. \tag{6.3}$$

Proof. Recall that for each a with $(a, p) = 1$, every integer x with $1 \leq x \leq p-1$ has a unique associate x' (with respect to a modulo p) of one another with $1 \leq x' \leq p-1$.

For quadratic residues a, we know from Theorem 6.2(i) that there are exactly two x's, say $x = x_1$ and $x = p - x_1$, whose associate is itself. Therefore, we may group $\{1, \ldots, p-1\}$ into $\{x_1\}$, $\{p-x_1\}$ and $\frac{p-3}{2}$ distinct unordered pairs $\{x, x'\}$ with

$$x_1^2 \equiv (p-x_1)^2 \equiv a \pmod{p}$$

and

$$xx' \equiv a \pmod{p}.$$

Thus,

$$(p-1)! = x_1 \cdot (p-x_1) \cdot \prod(xx') \equiv -x_1^2 \cdot \prod(xx') \equiv -a \cdot a^{\frac{p-3}{2}} = -a^{\frac{p-1}{2}} \pmod{p}.$$

For quadratic non-residues a, we cannot find any x such that $x^2 \equiv a \pmod{p}$. Therefore, we group $\{1,\ldots,p-1\}$ into $\frac{p-1}{2}$ distinct unordered pairs $\{x,x'\}$ with

$$xx' \equiv a \pmod{p}.$$

Thus,

$$(p-1)! = \prod(xx') \equiv a^{\frac{p-1}{2}} \pmod{p}.$$

The proof is therefore complete. ■

6.2 Wilson's Theorem

Let us take a look at the special case $a = 1$ of Theorem 6.3, which is known as *Wilson's Theorem*, named after the English mathematician John Wilson.

Theorem 6.4 (Wilson's Theorem). Let p be a prime. Then

$$(p-1)! \equiv -1 \pmod{p}. \tag{6.4}$$

Proof. If $p = 2$, we simply have $1 \equiv -1 \pmod{2}$, which is trivial. If p is an odd prime, then we note that 1 is a quadratic residue modulo p, for $1 \equiv 1^2 \pmod{p}$. Therefore, taking $a = 1$ in (6.2) yields (6.4). ■

Note that (6.4) is always false if the prime p is replaced with a composite.

Theorem 6.5 For $m \geq 2$, we have $(m-1)! \equiv -1 \pmod{m}$ if and only if m is prime.

Proof. The "if" part is exactly Wilson's Theorem. For the "only if" part, we assume that m is composite. Then m has a divisor d with $1 < d < m$. Thus, this d is among $2,\ldots,m-1$, and thus $d \mid (m-1)!$. This then implies that $d \nmid \big((m-1)!+1\big)$. But if $(m-1)! \equiv -1 \pmod{m}$, or equivalently, $m \mid \big((m-1)!+1\big)$, then all the divisors of m also divide $(m-1)!+1$, thereby leading to a contradiction. ■

6.3 Legendre symbol

We usually use the *Legendre symbol*, which was introduced by the French mathematician Adrien-Marie Legendre in 1798, to characterize whether an integer a is a quadratic residue modulo an odd prime p.

Definition 6.3 Let $p \geq 3$ be a prime and a be an integer. The *Legendre symbol* $\left(\frac{a}{p}\right)$ is defined by

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

Theorem 6.6 Let $p \geq 3$ be a prime, and a and b be such that $a \equiv b \pmod{p}$. Then

$$\left(\frac{a}{p}\right) = \left(\frac{b}{p}\right). \tag{6.5}$$

Proof. If $a \equiv b \equiv 0 \pmod{p}$, we have $\left(\frac{a}{p}\right) = \left(\frac{b}{p}\right) = 0$ by definition. If $a \equiv b \not\equiv 0 \pmod{p}$ and hence $(a,p) = (b,p) = 1$, the equality $\left(\frac{a}{p}\right) = \left(\frac{b}{p}\right)$ follows by noting that in this case, a and b are simultaneously quadratic residues modulo p, or simultaneously quadratic non-residues modulo p. ■

Theorem 6.7 Let $p \geq 3$ be a prime. Then

$$\sum_{a=1}^{p-1} \left(\frac{a}{p}\right) = 0. \tag{6.6}$$

In general, if $\{a_1, \ldots, a_{p-1}\}$ is a reduced system modulo p, then

$$\sum_{k=1}^{p-1} \left(\frac{a_k}{p}\right) = 0. \tag{6.7}$$

Proof. We simply make use of the fact from Theorem 6.2(ii) that there are exactly $\frac{p-1}{2}$ quadratic residue classes modulo p, and $\frac{p-1}{2}$ quadratic non-residue classes modulo p. ■

Theorem 6.8 Let $p \geq 3$ be a prime and a be such that $(a,p) = 1$. Then

$$\left(\frac{a}{p}\right) \equiv a^{\frac{p-1}{2}} \pmod{p}. \tag{6.8}$$

Proof. Note that Theorem 6.3 can be understood as

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

On the other hand, Wilson's Theorem asserts that

$$(p-1)! \equiv -1 \pmod{p}.$$

The desired result therefore follows. ■

Theorem 6.9 Let $p \geq 3$ be a prime, and m and n be integers. Then

$$\left(\frac{mn}{p}\right) = \left(\frac{m}{p}\right)\left(\frac{n}{p}\right). \tag{6.9}$$

Proof. If one of m and n is a multiple of p, so is mn. Thus, in this case,

$$\left(\frac{mn}{p}\right) = \left(\frac{m}{p}\right)\left(\frac{n}{p}\right) = 0.$$

Now we assume that $(m,p) = (n,p) = 1$ and thus $(mn,p) = 1$. Then by Theorem 6.8,

$$\left(\frac{mn}{p}\right) \equiv (mn)^{\frac{p-1}{2}} = m^{\frac{p-1}{2}} n^{\frac{p-1}{2}} \equiv \left(\frac{m}{p}\right)\left(\frac{n}{p}\right) \pmod{p},$$

that is, $p \mid \left|\left(\frac{mn}{p}\right) - \left(\frac{m}{p}\right)\left(\frac{n}{p}\right)\right|$. However, the values of $\left(\frac{m}{p}\right)$, $\left(\frac{n}{p}\right)$ and $\left(\frac{mn}{p}\right)$ are taken from $\{-1,1\}$. It follows that $\left|\left(\frac{mn}{p}\right) - \left(\frac{m}{p}\right)\left(\frac{n}{p}\right)\right| \le 2$, thereby implying that $\left(\frac{mn}{p}\right) - \left(\frac{m}{p}\right)\left(\frac{n}{p}\right) = 0$, as desired. ■

R Given a function $f : \mathbb{Z}_{>0} \to \mathbb{C}$, we say that it is *completely multiplicative* if $f(1) = 1$ and for any m and n,

$$f(mn) = f(m)f(n).$$

"Multiplicative" vs "Completely multiplicative": For completely multiplicative functions, the above relation holds even if $(m,n) > 1$.

6.4 When is -1 a quadratic residue modulo p?

Theorem 6.10 Let $p \ge 3$ be a prime. Then

$$\left(\frac{-1}{p}\right) = (-1)^{\frac{p-1}{2}}. \tag{6.10}$$

In particular, -1 is a quadratic residue modulo p if $p \equiv 1 \pmod 4$, and a quadratic non-residue modulo p if $p \equiv 3 \pmod 4$.

Proof. We know from Theorem 6.8 that $\left(\frac{-1}{p}\right) \equiv (-1)^{\frac{p-1}{2}} \pmod p$, and thus (6.10) follows since $\left(\frac{-1}{p}\right)$ takes value from $\{-1,1\}$ for odd primes p. Finally, $\frac{p-1}{2}$ is even if $p \equiv 1 \pmod 4$, and odd if $p \equiv 3 \pmod 4$. ■

6.5 Starters for sums of squares

We prove two additional results based on the knowledge of quadratic residues; they will be used in our later study of the "sum-of-squares" problems.

Theorem 6.11 Let $p \ge 3$ be a prime such that $p \equiv 1 \pmod 4$. Then there exists an integer x such that

$$x^2 + 1 = mp$$

with $0 < m < p$.

Proof. For primes $p \equiv 1 \pmod 4$, Theorem 6.10 tells us that -1 is a quadratic residue modulo p. Thus, there exists an x among $1, \ldots, p-1$ such that

$$x^2 \equiv -1 \pmod p.$$

In particular, we may choose x with $1 \le x \le \frac{p-1}{2}$, for if x satisfies the above congruence, so does $p - x$. Finally, we have $0 < x^2 + 1 < \left(\frac{p}{2}\right)^2 + 1 < p^2$. Thus, $x^2 + 1 = mp$ with $0 < m < p$. ■

Theorem 6.12 Let $p \ge 3$ be a prime. Then there exist integers x and y such that

$$x^2 + y^2 + 1 = mp$$

with $0 < m < p$.

Proof. Consider the following $p+1$ integers: x^2 for $0 \le x \le \frac{p-1}{2}$ and $-(y^2+1)$ for $0 \le y \le \frac{p-1}{2}$. Since there are p residue classes modulo p, by the pigeonhole principle, at least two of the $p+1$ integers fall into the same residue class. Note that all the x^2 are incongruent modulo p, and so are the $-(y^2+1)$. Thus, the two integers falling into the same residue class must be one x^2 and one $-(y^2+1)$. That is, there exist x and y with $0 \le x, y \le \frac{p-1}{2}$ such that $x^2 \equiv -(y^2+1) \pmod{p}$, namely, $x^2+y^2+1 = mp$ for an integer m. Finally, we have $0 < 1+x^2+y^2 < 1+2\left(\frac{p}{2}\right)^2 < p^2$. Thus, $0 < m < p$. ■

7. Quadratic reciprocity

7.1 Gauss's Lemma

To further evaluate the Legendre symbol $\left(\frac{a}{p}\right)$, we require a lemma due to the German mathematician Carl Friedrich Gauss. This lemma serves as a key in the derivation of the famous quadratic reciprocity law that was conjectured by Euler and Legendre, and first proved by Gauss himself.

Lemma 7.1 (Gauss's Lemma). Let $p \geq 3$ be a prime and a be such that $(a,p)=1$. For each k with $1 \leq k \leq \frac{p-1}{2}$, let r_k be the smallest nonnegative residue of ak modulo p. If $\mu = \mu_a$ counts the number of r_k that is greater than $\frac{p}{2}$, then

$$\left(\frac{a}{p}\right) = (-1)^{\mu}. \tag{7.1}$$

■ **Example 7.1** We provide an illustration of Gauss's Lemma with $p = 11$. Noting that the quadratic residues modulo 11 are given by the residue classes $\{1,3,4,5,9\}$ and that the non-residues are given by the residue classes $\{2,6,7,8,10\}$, we have, for instance, $\left(\frac{2}{11}\right) = -1$ and $\left(\frac{5}{11}\right) = 1$. **(i).** $a = 2$: We have $\{2k \bmod 11 : 1 \leq k \leq 5\} = \{2,4,\mathbf{6},\mathbf{8},\mathbf{10}\}$, and hence $\mu_2 = 3$ and $\left(\frac{2}{11}\right) = (-1)^3 = -1$; **(ii).** $a = 5$: We have $\{5k \bmod 11 : 1 \leq k \leq 5\} = \{5,\mathbf{10},4,\mathbf{9},3\}$, and hence $\mu_5 = 2$ and $\left(\frac{5}{11}\right) = (-1)^2 = 1$. ■

Proof. Since $(a,p)=1$, we have $1 \leq r_k \leq p-1$ for each k. Now we separate these indices $k \in \{1,2,\ldots,\frac{p-1}{2}\}$ into two disjoint groups $\{x_1,\ldots,x_\mu\}$ and $\{y_1,\ldots,y_\nu\}$ such that $r_x > \frac{p}{2}$ for all x and $r_y < \frac{p}{2}$ for all y. Note that $\mu + \nu = \frac{p-1}{2}$. Also, the r_x are pairwise distinct and so are the r_y. We further claim that there are no x and y with $p - r_x = r_y$; otherwise, we have $0 \equiv p = r_x + r_y \equiv ax + ay \pmod{p}$, or $x + y \equiv 0 \pmod{p}$, which is impossible since $1 \leq x,y \leq \frac{p-1}{2}$. As $1 \leq p - r_x < \frac{p}{2}$ and $1 \leq r_y < \frac{p}{2}$, we conclude that the $\frac{p-1}{2}$ pairwise distinct integers $(p - r_{x_1}),\ldots,(p - r_{x_\mu})$ and $r_{y_1},\ldots,r_{y_\nu}$ form a rearrangement of $1,\ldots,\frac{p-1}{2}$. Thus,

$$\begin{aligned} a^{\frac{p-1}{2}}\left(\frac{p-1}{2}\right)! &= \prod_{k=1}^{(p-1)/2}(ak) \equiv \prod_{k=1}^{(p-1)/2} r_k = \prod_{i=1}^{\mu} r_{x_i} \cdot \prod_{j=1}^{\nu} r_{y_j} \\ &\equiv (-1)^{\mu}\prod_{i=1}^{\mu}(p - r_{x_i}) \cdot \prod_{j=1}^{\nu} r_{y_j} = (-1)^{\mu}\left(\frac{p-1}{2}\right)! \pmod{p}. \end{aligned}$$

Since $(\frac{p-1}{2})!$ is coprime to p, we have $a^{\frac{p-1}{2}} \equiv (-1)^{\mu} \pmod{p}$. Finally, (7.1) follows since $\left(\frac{a}{p}\right)$ takes value from $\{\pm 1\}$ by definition and $\left(\frac{a}{p}\right) \equiv a^{\frac{p-1}{2}} \pmod{p}$ by Theorem 6.8. ■

For any real number x, let $\lfloor x \rfloor$ denote the largest integer not exceeding x.

Lemma 7.2 With the notation in Lemma 7.1, we have

$$\mu_a \equiv (a-1)\cdot\frac{p^2-1}{8} + \sum_{k=1}^{(p-1)/2}\left\lfloor \frac{ak}{p}\right\rfloor \pmod{2}. \tag{7.2}$$

Proof. Note that each r_k is the remainder of ak divided by p. Thus, $ak = p\cdot\lfloor\frac{ak}{p}\rfloor + r_k$. Now, recalling that p is an odd prime,

$$\begin{aligned} a\cdot\frac{p^2-1}{8} &= \sum_{k=1}^{(p-1)/2}(ak) = \sum_{k=1}^{(p-1)/2}\left(p\cdot\left\lfloor\frac{ak}{p}\right\rfloor + r_k\right) = p\sum_{k=1}^{(p-1)/2}\left\lfloor\frac{ak}{p}\right\rfloor + \sum_{i=1}^{\mu} r_{x_i} + \sum_{j=1}^{\nu} r_{y_j} \\ &\equiv \sum_{k=1}^{(p-1)/2}\left\lfloor\frac{ak}{p}\right\rfloor + \left(\mu + \sum_{i=1}^{\mu}(p - r_{x_i})\right) + \sum_{j=1}^{\nu} r_{y_j} = \sum_{k=1}^{(p-1)/2}\left\lfloor\frac{ak}{p}\right\rfloor + \mu + \sum_{k=1}^{(p-1)/2} k \\ &= \sum_{k=1}^{(p-1)/2}\left\lfloor\frac{ak}{p}\right\rfloor + \mu + \frac{p^2-1}{8} \pmod{2}, \end{aligned}$$

thereby yielding the desired result. ■

7.2 When is 2 a quadratic residue modulo p?

Theorem 7.3 Let $p \geq 3$ be a prime. Then

$$\left(\frac{2}{p}\right) = (-1)^{\frac{p^2-1}{8}}. \tag{7.3}$$

In particular, 2 is a quadratic residue modulo p if $p \equiv \pm 1 \pmod{8}$, and a quadratic non-residue modulo p if $p \equiv \pm 3 \pmod{8}$.

Proof. Note that for k with $1 \leq k \leq \frac{p-1}{2}$, we have $0 < \frac{2k}{p} < 1$ and thus $\lfloor\frac{2k}{p}\rfloor = 0$. Now, taking $a = 2$ in (7.2) gives $\mu_2 \equiv \frac{p^2-1}{8} \pmod{2}$, and it follows from Gauss's Lemma that $\left(\frac{2}{p}\right) = (-1)^{\frac{p^2-1}{8}}$. Finally, $\frac{p^2-1}{8}$ is even if $p \equiv \pm 1 \pmod{8}$, and odd if $p \equiv \pm 3 \pmod{8}$. ■

7.3 Guass's law of quadratic reciprocity

We have witnessed from Gauss's Lemma (Lemma 7.1) and Lemma 7.2 that for $p \geq 3$ a prime and a an integer with $(a,p) = 1$,

$$\left(\frac{a}{p}\right) = (-1)^{(a-1)\cdot\frac{p^2-1}{8} + \sum_{k=1}^{(p-1)/2}\lfloor\frac{ak}{p}\rfloor}.$$

Now we further assume that $q \geq 3$ is a prime such that $q \neq p$. Then $(q-1)\cdot\frac{p^2-1}{8}$ is even for $q-1$ is even and $\frac{p^2-1}{8} = \sum_{k=1}^{(p-1)/2} k$ is an integer. It follows that

$$\left(\frac{q}{p}\right) = (-1)^{\sum_{k=1}^{(p-1)/2}\lfloor\frac{kq}{p}\rfloor}.$$

Similarly,

$$\left(\frac{p}{q}\right) = (-1)^{\sum_{k=1}^{(q-1)/2} \lfloor \frac{kp}{q} \rfloor}.$$

It turns out that

$$\left(\frac{q}{p}\right)\left(\frac{p}{q}\right) = (-1)^{\sum_{k=1}^{(p-1)/2} \lfloor \frac{kq}{p} \rfloor + \sum_{k=1}^{(q-1)/2} \lfloor \frac{kp}{q} \rfloor}. \tag{7.4}$$

Theorem 7.4 Let $p, q \geq 3$ be primes with $p \neq q$. Then

$$\sum_{k=1}^{(p-1)/2} \left\lfloor \frac{kq}{p} \right\rfloor + \sum_{k=1}^{(q-1)/2} \left\lfloor \frac{kp}{q} \right\rfloor = \frac{(p-1)(q-1)}{4}. \tag{7.5}$$

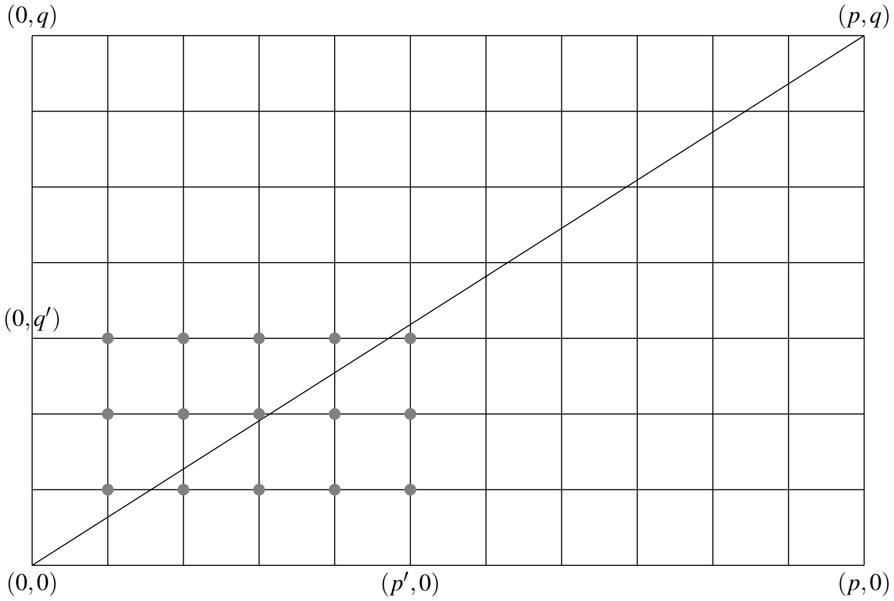


Figure 7.1: Integer lattices and $y = \frac{q}{p}x$

Proof. For convenience, we write $p' = \frac{p-1}{2}$ and $q' = \frac{q-1}{2}$. Consider the line

$$\ell : y = \frac{q}{p}x$$

on the xy-plane. We begin with some observations.

Observation 1. *For any integer $k \geq 1$, $\lfloor \frac{kq}{p} \rfloor$ equals the number of points with integer coordinates, or lattices for short, (k, y) in the first quadrant which are below ℓ (with lattices on ℓ included).* For its proof, we note that ℓ touches the vertical line $x = k$ at $(k, \frac{kq}{p})$. Thus, such lattices are those with $1 \leq y \leq \frac{kq}{p}$, and the number of them equals the integer part of $\frac{kq}{p}$, that is $\lfloor \frac{kq}{p} \rfloor$.

Observation 2. *For any integer $k \geq 1$, $\lfloor \frac{kp}{q} \rfloor$ equals the number of lattices (x, k) in the first quadrant which are above ℓ (with lattices on ℓ included).* The proof is similar to that for the first observation — we only need to note that ℓ touches the horizontal line $y = k$ at $(\frac{kp}{q}, k)$.

Observation 3. *There is no lattice* (x,y) *with* $1 \le x \le p'$ *or* $1 \le y \le q'$ *that is on* ℓ*.* Otherwise, assume that there exists an x_0 with $1 \le x_0 \le p'$ such that $(x_0, \frac{q}{p}x_0)$ is a lattice. Then $\frac{q}{p}x_0$ is an integer, which is impossible since $p \nmid q$ for p,q are distinct odd primes and $p \nmid x_0$ for $1 \le x \le p' = \frac{p-1}{2}$. Similarly, if we assume that there exists a y_0 with $1 \le y_0 \le q'$ such that $(\frac{p}{q}y_0, y_0)$ is a lattice, then $\frac{p}{q}y_0$ is an integer, and it is also impossible. The claim follows by contradiction.

Now, we focus on the set of lattices (x,y) with $1 \le x \le p'$ and $y \ge 1$ that are **strictly** below ℓ, denoted by $\mathscr{B}$, and the set of lattices (x,y) with $x \ge 1$ and $1 \le y \le q'$ that are **strictly** above ℓ, denoted by $\mathscr{A}$.

By the three observations (especially Observation 3, which allows us to add the strengthening of "**strictly**"), we have

$$\sum_{k=1}^{(p-1)/2} \left\lfloor \frac{kq}{p} \right\rfloor + \sum_{k=1}^{(q-1)/2} \left\lfloor \frac{kp}{q} \right\rfloor = \operatorname{card} \mathscr{A} + \operatorname{card} \mathscr{B}.$$

First, it is apparent that all lattices (x,y) with $1 \le x \le p'$ **and** $1 \le y \le q'$ are in $\mathscr{A} \cup \mathscr{B}$. Now we show that they are the only lattices in $\mathscr{A} \cup \mathscr{B}$.

(i). For lattices with $x > p'$ and $y > q'$, they are not in $\mathscr{A} \cup \mathscr{B}$ by definition.

(ii). For any lattice with $1 \le x \le p'$ and $y > q'$ (so it is not in $\mathscr{A}$), we compute the slope of the line connecting this lattice and the origin, which is $\frac{y}{x} \ge \frac{q'+1}{p'} = \frac{q+1}{p-1} > \frac{q}{p}$, and thus the lattice is above ℓ, so not in $\mathscr{B}$.

(iii). For any lattice with $x > p'$ and $1 \le y \le q'$ (so it is not in $\mathscr{B}$), we compute the slope of the line connecting this lattice and the origin, which is $\frac{y}{x} \le \frac{q'}{p'+1} = \frac{q-1}{p+1} < \frac{q}{p}$, and thus the lattice is below ℓ, so not in $\mathscr{A}$.

Noting that $\mathscr{A}$ and $\mathscr{B}$ are disjoint, we have $\operatorname{card} \mathscr{A} + \operatorname{card} \mathscr{B} = \operatorname{card} \mathscr{A} \cup \mathscr{B} = p'q'$. Thus,

$$\sum_{k=1}^{(p-1)/2} \left\lfloor \frac{kq}{p} \right\rfloor + \sum_{k=1}^{(q-1)/2} \left\lfloor \frac{kp}{q} \right\rfloor = \operatorname{card} \mathscr{A} + \operatorname{card} \mathscr{B} = p'q' = \frac{(p-1)(q-1)}{4},$$

thereby proving the desired result. ■

Now, we can state *Guass's law of quadratic reciprocity.*

Theorem 7.5 (Guass's Law of Quadratic Reciprocity). Let $p,q \ge 3$ be primes with $p \ne q$. Then

$$\left(\frac{q}{p}\right)\left(\frac{p}{q}\right) = (-1)^{\frac{(p-1)(q-1)}{4}}. \tag{7.6}$$

Proof. This is a direct application of (7.4) and (7.5). ■

7.4 When is ± 3 a quadratic residue modulo p?

Theorem 7.6 Let $p \ge 5$ be a prime. Then 3 is a quadratic residue modulo p if $p \equiv \pm 1 \pmod{12}$, and a quadratic non-residue modulo p if $p \equiv \pm 5 \pmod{12}$.

Proof. By Guass's law of quadratic reciprocity, we have

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

Further, $\left(\frac{p}{3}\right)$ equals 1 if $p \equiv 1 \pmod 3$ and -1 if $p \equiv -1 \pmod 3$. Also, $(-1)^{\frac{p-1}{2}}$ equals 1 if $p \equiv 1 \pmod 4$ and -1 if $p \equiv -1 \pmod 4$. The desired result follows from a simple calculation. ■

Theorem 7.7 Let $p \geq 5$ be a prime. Then -3 is a quadratic residue modulo p if $p \equiv 1 \pmod 6$, and a quadratic non-residue modulo p if $p \equiv 5 \pmod 6$.

Proof. Note that

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

Combining Theorems 6.10 and 7.6 gives the desired result. ■

7.5 An upper bound for the least quadratic non-residue

Definition 7.1 Let $p \geq 3$ be a prime. The *least quadratic non-residue modulo* p, usually denoted by n_p, is the smallest positive integer that is a quadratic non-residue modulo p.

Example 7.2 We have $n_3 = 2$, $n_5 = 2$, $n_7 = 3$, ...

R The least quadratic residue is less interesting because 1 is always a quadratic residue modulo any odd prime p.

Recall from Theorem 6.2 that there are $\frac{p-1}{2}$ quadratic residues and $\frac{p-1}{2}$ quadratic non-residues modulo p among $1,\dots,p-1$. Therefore, we trivially have $n_p \leq \frac{p-1}{2} + 1 = \frac{p+1}{2}$. But the upper bound for n_p could be much sharper.

Theorem 7.8 Let $p \geq 3$ be a prime. Then

$$n_p < \sqrt{p} + 1. \tag{7.7}$$

Proof. Note that $1 < n_p < p$. Let $m = \lfloor \frac{p}{n_p} \rfloor + 1$. Since $\frac{p}{n_p}$ is not an integer, we have $(m-1)n_p < p < mn_p$. Thus, $0 < mn_p - p < n_p$. Since n_p is the least quadratic non-residue, we have that all $1,\dots,n_p - 1$ are quadratic residues, and so is $mn_p - p$. It follows that

$$1 = \left(\frac{mn_p - p}{p}\right) = \left(\frac{mn_p}{p}\right) = \left(\frac{m}{p}\right)\left(\frac{n_p}{p}\right),$$

where Theorem 6.9 is used. Since n_p is a quadratic non-residue, we have $\left(\frac{n_p}{p}\right) = -1$, and thus $\left(\frac{m}{p}\right) = -1$ from the above. Thus, m is also a quadratic non-residue. It follows that $n_p \leq m$. So,

$$p > (m-1)n_p \geq (n_p - 1)n_p > (n_p - 1)^2,$$

yielding the desired result. ■

R The upper bound for n_p is far sharper than (7.7). The best bound known today is

$$n_p = O_\varepsilon\left(p^{\frac{1}{4\sqrt{e}}+\varepsilon}\right),$$

for all $\varepsilon > 0$. It was proved with recourse to Burgess's estimate of certain character sums and Vinogradov's sieving trick. An excellent exposition of the idea can be found in Terry Tao's blog post:

https://terrytao.wordpress.com/2009/08/18/the-least-quadratic-nonresidue-and-the-square-root-barrier/

8. Sums of squares

8.1 Primes as the sum of two squares

Recall that the following notation has been used earlier in the study of binomial coefficients.

Definition 8.1 Let p be a prime. Given any nonzero integer n, we denote by $\nu_p(n)$ the unique nonnegative integer k such that $p^k \mid n$ and $p^{k+1} \nmid n$, namely, $\nu_p(n)$ is the power of p in the canonical form of n.

Theorem 8.1 Let x and y be integers, not both zero. For any prime p with $p \equiv 3 \pmod 4$, we have that $\nu_p(x^2+y^2)$ is even.

Proof. Let $n = x^2+y^2$. Note that $n > 0$. Let $d = (x,y)$ and write $x = x_0 d$ and $y = y_0 d$ so $(x_0,y_0) = 1$. Hence, $n = d^2(x_0^2+y_0^2)$.

We first show that $p \nmid (x_0^2+y_0^2)$. If not, then $x_0^2+y_0^2 \equiv 0 \pmod p$, or $x_0^2 \equiv -y_0^2 \pmod p$. Since $(x_0,y_0) = 1$, both x_0 and y_0 are coprime to p for if any of them is a multiple of p, so is the other. Now y_0 has an inverse y_0^{-1} modulo p. Hence, $(x_0y_0^{-1})^2 \equiv -1 \pmod p$, indicating that -1 is a quadratic residue modulo p. However, this violates Theorem 6.10, saying that -1 is a quadratic non-residue as $p \equiv 3 \pmod 4$.

Thus, $\nu_p(n) = \nu_p(d^2) = 2\nu_p(d)$, which is even. ■

Theorem 8.2 Any prime p with $p \equiv 1 \pmod 4$ can be written as the sum of two squares.

We will present two proofs of this result: one is based on an important method called "infinite descent" developed by Fermat, and the other relies on a magical involution due to the American–German mathematician Don Zagier.

Before moving ahead, we record a simple but useful formula.

Lemma 8.3 Let $x_1,y_1,x_2,y_2 \in \mathbb{R}$. Then

$$(x_1^2+y_1^2)(x_2^2+y_2^2) = (x_1x_2+y_1y_2)^2+(x_1y_2-x_2y_1)^2. \tag{8.1}$$

Proof. This formula can be examined by a direct calculation. ■

R We may also understand (8.1) with recourse to complex numbers. Recall that a *complex number* z is of the form $z=x+yi$ with $x,y\in\mathbb{R}$ where $i=\sqrt{-1}$ is the *imaginary unit*. The *modulus* of z is define by $|z|=\sqrt{x^2+y^2}$. Let $z_1=x_1-y_1i$ and $z_2=x_2+y_2i$. Note that the left hand side of (8.1) is $|z_1|^2|z_2|^2$ and the right hand side is $|z_1z_2|^2$. So, $|z_1|^2|z_2|^2=|z_1z_2|^2$.

8.2 The method of infinite descent

Among different variants of the method of *infinite descent*, which is also known as *Fermat's method of descent* as Fermat developed this strategy, which first appeared in Euclid's *Elements*, to a great extent, we will make use of the following version.

Lemma 8.4 (The Method of Infinite Descent). Let P be a property that at least one positive integer, say M, possesses. Assume that whenever a positive integer m with $1<m\le M$ possesses P, we may find another positive integer m_0 with $m_0<m$ such that m_0 also possesses P. Then 1 possesses P.

Proof. We argue by contradiction with the assumption that 1 does not possess P. Since P is such that M possesses, we may let $m'\le M$ be the smallest positive integer possessing P. By our assumption, $m'>1$. However, we may then find some $m_0'<m'$ with m_0' possessing P. This violates the minimality of m'. ■

Now we prove Theorem 8.2 using the method of infinite descent.

First Proof of Theorem 8.2. To begin with, we recall from Theorem 6.11 that for primes p with $p\equiv 1 \pmod 4$, there exists an integer x such that $x^2+1=mp$ with $0<m<p$. In other words, there exists an integer m with $0<m<p$ such that the equation

$$x^2+y^2=mp$$

has an integer solution (x,y).

Assume that $m>1$. Note that for every integer n, we may always find an integer n_0 with $|n_0|\le\frac{m}{2}$ such that $n\equiv n_0 \pmod m$. This is because there are at least m consecutive integers in the interval $[-\frac{m}{2},\frac{m}{2}]$, thereby covering a complete system modulo m.

Now we find $x\equiv x_0 \pmod m$ with $|x_0|\le\frac{m}{2}$ and $y\equiv y_0 \pmod m$ with $|y_0|\le\frac{m}{2}$. Note that we cannot simultaneously have $m\mid x$ and $m\mid y$ for if this is the case, then it follows that $m^2\mid(x^2+y^2)=mp$. But $m^2\nmid mp$ since $0<m<p$ (and hence $(m,p)=1$), thereby leading to a contradiction. Hence, x_0 and y_0 are not simultaneously 0, and we have $x_0^2+y_0^2>0$. On the other hand, $x_0^2+y_0^2\le(\frac{m}{2})^2+(\frac{m}{2})^2<m^2$. Noting that $x_0^2+y_0^2\equiv x^2+y^2=mp\equiv 0 \pmod m$, we may write $x_0^2+y_0^2=m_0m$ with $0<m_0<m$. By Lemma 8.3, we have

$$\begin{aligned}(xx_0+yy_0)^2+(xy_0-x_0y)^2&=(x^2+y^2)(x_0^2+y_0^2)\\&=(mp)\cdot(m_0m)\\&=m^2m_0p.\end{aligned}$$

Meanwhile, we have $xx_0+yy_0\equiv x^2+y^2\equiv 0 \pmod m$ and $xy_0-x_0y\equiv xy-xy=0 \pmod m$. Hence, $\frac{xx_0+yy_0}{m}$ and $\frac{xy_0-x_0y}{m}$ are integers. It follows that

$$m_0p=\left(\frac{xx_0+yy_0}{m}\right)^2+\left(\frac{xy_0-x_0y}{m}\right)^2,$$

a sum of two squares.

Finally, noting that m_0 is a positive integer with $m_0<m$, we deduce that $x^2+y^2=p$ has an integer solution (x,y) with recourse to the method of infinite descent. ■

8.3 Zagier's magical involution

Definition 8.2 Let S be a set. We say that $f : S \to S$ is an *involution* on S if for any $x \in S$, there holds true that $f(f(x)) = x$.

R In fact, every involution f is a bijective map on S. The surjectivity follows by the fact every $x \in S$ in the image of $f(x)$ under f, and the injectivity follows by the fact that if $f(x) = f(y)$, then $x = f(f(x)) = f(f(y)) = y$.

Definition 8.3 Let S be a set and $f : S \to S$ be a map on S. We say that $x \in S$ is a *fixed point* under f if $f(x) = x$.

Theorem 8.5 Let S be a finite set and assume that there is an involution f on S.

(i) If f has no fixed points, then the size $|S|$ of S is even.

(ii) If f has exactly one fixed point, then $|S|$ is odd.

Proof. Since f is an involution on S, we may pair elements of S according to $\{x, f(x)\}$ and treat $\{f(x), x\}$ as the same pair. Assume that there are s such pairs.

(i). Since f has no fixed points, we have $x \neq f(x)$ in each pair. Thus, every $x \in S$ belongs to exactly one of the pairs. It follows that $|S| = 2s$, which is even.

(ii). Assume that the only fixed point of f is x_0. Every $x \in S$ is either x_0, or belongs to exactly one of the pairs, excluding $\{x_0, f(x_0)\} = \{x_0, x_0\}$. Thus, $|S| = 1 + 2(s-1) = 2s-1$, which is odd. ■

Theorem 8.6 Let p be a prime with $p \equiv 1 \pmod 4$. Consider the finite set $S = \{(x,y,z) \in \mathbb{Z}^3_{>0} : x^2 + 4yz = p\}$. Then the following map f on S,

$$f(x,y,z) = \begin{cases} (x+2z, z, y-x-z), & \text{if } x < y-z, \\ (2y-x, y, x-y+z), & \text{if } y-z < x < 2y, \\ (x-2y, x-y+z, y), & \text{if } x > 2y, \end{cases}$$

is an involution, and it has exactly one fixed point. In particular, $|S|$ is odd.

Proof. We first show that $x \neq y-z$ and $x \neq 2y$ for $(x,y,z) \in S$. If $x = y-z$, then $p = (y-z)^2 + 4yz = (y+z)^2$, which is impossible since p is prime. If $x = 2y$, then $p = (2y)^2 + 4yz = 4y(y+z)$, which is also impossible. Thus, we may separate S into three disjoint subsets S_1, S_2 and S_3 according to **(1).** $x < y-z$, **(2).** $y-z < x < 2y$, **(3).** $x > 2y$.

A direct calculation reveals that for any $(x,y,z) \in S$, $f(f(x,y,z)) = (x,y,z)$, and hence, f is an involution. Also, if $(x,y,z) \in S_1$, then $f(x,y,z) \in S_3$; if $(x,y,z) \in S_2$, then $f(x,y,z) \in S_2$; and if $(x,y,z) \in S_3$, then $f(x,y,z) \in S_1$. Hence, fixed points (x,y,z) are only in S_2, with

$$x = 2y - x, \qquad y = y, \qquad z = x - y + z,$$

namely, $x = y$. But in this case, $p = x^2 + 4xz = x(x+4z)$ implies that the only possible x is $x = 1$, and hence $y = x = 1$. Finally, since $p \equiv 1 \pmod 4$, that is, $p = 4k+1$ with $k > 0$, we have the unique fixed point $(x,y,z) = (1,1,k)$. We conclude from Theorem 8.5 that $|S|$ is odd. ■

Theorem 8.6 immediately yields an alternative proof of Theorem 8.2.

Second Proof of Theorem 8.2. The set S in Theorem 8.6 also has a trivial involution g given by $g(x,y,z) = (x,z,y)$. But g must have a fixed point; otherwise, $|S|$ is even by Theorem 8.5, thereby contradicting Theorem 8.6. But the fixed point of g means that $z = y$. Hence, we may find positive integers x and y such that $p = x^2 + 4y^2 = x^2 + (2y)^2$. ■

R Don Zagier's proof was published in (*Amer. Math. Monthly* **97** (1990), no. 2, 144). In fact, his involution is a refinement of an equally beautiful argument attributed to Roger Heath-Brown (*Invariant* (1984), 2–5). Heath-Brown's proof, dating back to 1971, was motivated by his study of J. V. Uspensky and M. A. Heaslet's book "*Elementary Number Theory*" (McGraw-Hill Book Co., Inc., New York, 1939), which accounts Liouville's papers on identities for parity functions.

8.4 Fermat's two-square theorem

Now we are in a position to characterize which integers can be written as the sum of two squares.

Theorem 8.7 (Fermat's Two-Square Theorem). A positive integer n can be written as the sum of two squares if and only if all prime factors p of n with $p \equiv 3 \pmod 4$ have an even power in the canonical form of n.

Proof. The "only if" part has been shown by Theorem 8.1. For the "if" part, we write in the canonical form

$$n = 2^{\alpha} \prod_{p \equiv 1 \bmod 4} p^{\beta} \prod_{q \equiv 3 \bmod 4} q^{2\gamma}.$$

Here, p runs over all distinct prime factors of n that are congruent to 1 modulo 4, and q runs over all distinct prime factors of n that are congruent to 3 modulo 4. In particular, the exponent of each q is even as assumed. Now, note that $2 = 1^2 + 1^2$, that $q^2 = 0^2 + q^2$ for each q, and that $p = x^2 + y^2$ for certain integers x and y by Theorem 8.2 for each p. A repeated application of Lemma 8.3 gives the desired result. ■

8.5 Lagrange's four-square theorem

Concerning sums of four squares, we first require an analog of Lemma 8.3.

Lemma 8.8 Let $x_1, y_1, z_1, w_1, x_2, y_2, z_2, w_2 \in \mathbb{R}$. Then

$$\begin{aligned}(x_1^2 + y_1^2 + z_1^2 + w_1^2)&(x_2^2 + y_2^2 + z_2^2 + w_2^2)\\ &= (x_1x_2 + y_1y_2 + z_1z_2 + w_1w_2)^2 + (x_1y_2 - y_1x_2 + z_1w_2 - w_1z_2)^2\\ &\quad + (x_1z_2 - y_1w_2 - z_1x_2 + w_1y_2)^2 + (x_1w_2 + y_1z_2 - z_1y_2 - w_1x_2)^2. \qquad (8.2)\end{aligned}$$

Proof. This formula can also be examined by a direct calculation. ■

Theorem 8.9 (Lagrange's Four-Square Theorem). Every positive integer can be written as the sum of four squares.

Proof. Note that $1 = 0^2 + 0^2 + 0^2 + 1^2$ and $2 = 0^2 + 0^2 + 1^2 + 1^2$. In view of Lemma 8.8, it suffices to show that every odd prime can be written as the sum of four squares.

Recall from Theorem 6.12 that for odd primes p, there exist integers x and y such that $x^2+y^2+1=mp$ with $0<m<p$. In other words, there exists an integer m with $0<m<p$ such that the equation

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

has an integer solution (x,y,z,w).

Assume that $m>1$. We have two cases.

(i). If m is even, then two of the integers x, y, z and w have the same parity, and the remaining two also have the same parity. Without loss of generality, we assume that x and y have the same parity, and z and w have the same parity. Thus, the four integers $x+y$, $x-y$, $z+w$, $z-w$ are even. Note that if $m_0=\frac{m}{2}$, then $0<m_0<m$. Also,

$$\begin{aligned} m_0p &= \frac{1}{2}(x^2+y^2+z^2+w^2) \\ &= \left(\frac{x+y}{2}\right)^2+\left(\frac{x-y}{2}\right)^2+\left(\frac{z+w}{2}\right)^2+\left(\frac{z-w}{2}\right)^2, \end{aligned}$$

a sum of four squares.

(ii). If m is odd, then similar to the first proof of Theorem 8.2, we find $x\equiv x_0 \pmod m$ with $|x_0|<\frac{m}{2}$, $y\equiv y_0 \pmod m$ with $|y_0|<\frac{m}{2}$, $z\equiv z_0 \pmod m$ with $|z_0|<\frac{m}{2}$ and $w\equiv w_0$ $(\text{mod } m)$ with $|w_0|<\frac{m}{2}$. Here, we use strict "$<$" since m is odd. Therefore, $x_0^2+y_0^2+z_0^2+w_0^2<(\frac{m}{2})^2+(\frac{m}{2})^2+(\frac{m}{2})^2+(\frac{m}{2})^2=m^2$. Also, we cannot simultaneously have $m\mid x$, $m\mid y$, $m\mid z$ and $m\mid w$, and hence, $x_0^2+y_0^2+z_0^2+w_0^2>0$. Noting that $x_0^2+y_0^2+z_0^2+w_0^2\equiv x^2+y^2+z^2+w^2=mp\equiv 0 \pmod m$, we may write $x_0^2+y_0^2+z_0^2+w_0^2=m_0m$ with $0<m_0<m$. By Lemma 8.8,

$$\begin{aligned} m^2m_0p &= (mp)\cdot(m_0m) = (x^2+y^2+z^2+w^2)(x_0^2+y_0^2+z_0^2+w_0^2) \\ &= (xx_0+yy_0+zz_0+ww_0)^2+(xy_0-yx_0+zw_0-wz_0)^2 \\ &\quad +(xz_0-yw_0-zx_0+wy_0)^2+(xw_0+yz_0-zy_0-wx_0)^2 \\ &=: \tilde{x}^2+\tilde{y}^2+\tilde{z}^2+\tilde{w}^2. \end{aligned}$$

Since $x\equiv x_0 \pmod m$, $y\equiv y_0 \pmod m$, $z\equiv z_0 \pmod m$, $w\equiv w_0 \pmod m$ and $x^2+y^2+z^2+w^2\equiv 0 \pmod m$, we find that $\tilde{x}$, $\tilde{y}$, $\tilde{z}$ and $\tilde{w}$ are all multiples of m. Hence,

$$m_0p=\left(\frac{\tilde{x}}{m}\right)^2+\left(\frac{\tilde{y}}{m}\right)^2+\left(\frac{\tilde{z}}{m}\right)^2+\left(\frac{\tilde{w}}{m}\right)^2,$$

a sum of two squares.

Finally, noting that in both cases of the above, m_0 is a positive integer with $m_0<m$, we deduce that $x^2+y^2+z^2+w^2=p$ has an integer solution (x,y,z,w) with recourse to the method of infinite descent. ■

9. Generating functions

9.1 Generating functions

In the previous lecture, we have shown the existence of a representation as the sum of four squares for each nonnegative integer n. Now a natural question is how many such representations do we have? Is there a formula, or at least a nice way, to characterize the number of such representations for each n?

In general, for $\{a_n\}_{n\geq 0}$ a sequence of numbers, not necessarily integers, we want to find a clothesline on which we hang up $\{a_n\}$ for display.

Definition 9.1 Let $\{a_n\}_{n\geq 0}$ be a sequence of numbers. Then the power series

$$\sum_{n\geq 0} a_n x^n = a_0 + a_1 x + a_2 x^2 + \cdots$$

is called the *generating function* of $\{a_n\}$.

R Since we are working on infinite series, a natural question is their radii of convergence. However, this question is usually not very interesting for generating functions, and in many cases we only treat these power series in a *formal* way. Nonetheless there are still occasions on which the radii of convergence should be taken into account, especially when analytic techniques are applied. For instance, when we want to make use of *Cauchy's integral formula* to recover the coefficients a_n from its generating function $A(x) = \sum_{n\geq 0} a_n x^n$:

$$a_n = \frac{1}{2\pi i} \oint \frac{A(x)}{x^{n+1}} dx,$$

we must be careful about the convergence conditions when choosing the contour.

9.2 Formal power series

Definition 9.2 A *formal power series* is an expression of the form

$$a_0 + a_1 x + a_2 x^2 + \cdots,$$

where the sequence $\{a_n\}_{n\geq 0}$ is called the *sequence of coefficients*.

We say two series $A(x) = \sum_{n\geq 0} a_n x^n$ and $B(x) = \sum_{n\geq 0} b_n x^n$ are *equal* if $a_n = b_n$ for all $n \geq 0$. We can also define the usual operations for formal power series:

▷ *Addition/Subtraction*:

$$\sum_{n\geq 0} a_n x^n \pm \sum_{n\geq 0} b_n x^n = \sum_{n\geq 0} (a_n \pm b_n) x^n;$$

▷ *Multiplication* by the Cauchy product rule:

$$\left(\sum_{n\geq 0} a_n x^n\right)\left(\sum_{n\geq 0} b_n x^n\right) = \sum_{n\geq 0} c_n x^n, \quad \text{where } c_n = \sum_{k=0}^{n} a_k b_{n-k}.$$

To determine if division works, we need to check if a series has a *reciprocal.*

Definition 9.3 Given a formal power series $\sum_{n\geq 0} a_n x^n$, we say a series $\sum_{n\geq 0} b_n x^n$ is the *reciprocal* of $\sum_{n\geq 0} a_n x^n$ if

$$\left(\sum_{n\geq 0} a_n x^n\right)\left(\sum_{n\geq 0} b_n x^n\right) = 1.$$

Theorem 9.1 A formal power series $A(x) = \sum_{n\geq 0} a_n x^n$ has a reciprocal if and only if $a_0 \neq 0$. In this case, the reciprocal is unique.

Proof. **(i).** If $A(x)$ has a reciprocal, say $B(x) = \sum_{n\geq 0} b_n x^n$. Then $A(x)B(x) = 1$. Hence, $a_0 b_0 = 1$, which implies that $a_0 \neq 0$. Further, b_0 is uniquely given by $1/a_0$. Also, for $n \geq 1$, we have $0 = \sum_{k=0}^{n} a_k b_{n-k}$. Therefore,

$$b_n = -\frac{1}{a_0} \sum_{k=1}^{n} a_k b_{n-k}.$$

By induction, the b_n's are uniquely determined.

(ii). If $a_n \neq 0$, we choose $b_0 = 1/a_0$, and iteratively define $b_n = -\frac{1}{a_0}\sum_{k=1}^{n} a_k b_{n-k}$. Then we get a series $B(x) = \sum_{n\geq 0} b_n x^n$. It is straightforward to verify that $A(x)B(x) = 1$, and hence $B(x)$ is a reciprocal of $A(x)$. ■

▪ **Example 9.1** We have

$$(1-x)(1+x+x^2+\cdots) = 1.$$

Hence, the reciprocal of $1-x$ is given by $1+x+x^2+\cdots$, written as

$$\frac{1}{1-x} = 1+x+x^2+\cdots.$$

This is exactly identical to what is obtained by applying the Taylor expansion to $\frac{1}{1-x}$. ▪

Definition 9.4 Let $A(x) = \sum_{n\geq 0} a_n x^n$ be a formal power series. Its *derivative* is the series

$$A'(x) = \sum_{n\geq 1} n a_n x^{n-1}.$$

▪ **Example 9.2** We know that

$$e^x = \sum_{n\geq 0} \frac{x^n}{n!}.$$

Now,

$$\left(\sum_{n\geq 0} \frac{x^n}{n!}\right)' = \sum_{n\geq 1} \frac{n x^{n-1}}{n!} = \sum_{n\geq 1} \frac{x^{n-1}}{(n-1)!} = e^x.$$

This is exactly identical to $(e^x)' = e^x$. ▪

9.3 Fibonacci numbers

The *Fibonacci numbers* are named after the Italian mathematician Leonardo of Pisa, later known as Fibonacci, for his famous "*Rabbit Puzzle*" in his 1202 book *Liber Abaci*:

> *Assume that we have a pair of fictional rabbits:*
>
> (i) *they produce a new pair of rabbits every month, starting from the second month that they are alive;*
>
> (ii) *and the new generations always repeat the trajectory of their parents' life.*
>
> *If rabbits never die and continue breeding forever, how many pairs will there be in one year?*

Assume that there are F_n pairs of rabbits after n months, starting with $F_0 = 0$ and $F_1 = 1$. Now, for F_n with $n \geq 2$, the rabbits are from the alive ones of the previous month, F_{n-1} pairs in total, and the newly born rabbits produced by those of at least two-month-old, F_{n-2} pairs in total. Therefore, for $n \geq 2$,

$$F_n = F_{n-1} + F_{n-2}. \tag{9.1}$$

Theorem 9.2 We have

$$\sum_{n\geq 0} F_n x^n = \frac{x}{1-x-x^2}. \tag{9.2}$$

Proof. We multiply (9.1) by x^n, and sum over $n \geq 2$. Then

$$\sum_{n\geq 2} F_n x^n = \sum_{n\geq 2} (F_{n-1}+F_{n-2})x^n = x\sum_{n\geq 2} F_{n-1}x^{n-1} + x^2 \sum_{n\geq 2} F_{n-2}x^{n-2} = x\sum_{n\geq 1} F_n x^n + x^2 \sum_{n\geq 0} F_n x^n.$$

Let $f(x) = \sum_{n\geq 0} F_n x^n$. We have

$$f(x) - (0+x) = x\big(f(x)-0\big) + x^2 f(x),$$

or

$$(1-x-x^2)f(x) = x.$$

This gives the desired result. ■

Can we find an explicit formula for F_n?

Theorem 9.3 For $n \geq 0$,

$$F_n = \frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right) \tag{9.3}$$

Proof. Let $\alpha = \frac{1+\sqrt{5}}{2}$ and $\beta = \frac{1-\sqrt{5}}{2}$. Then $1-x-x^2 = (1-\alpha x)(1-\beta x)$. Therefore,

$$\begin{aligned}\frac{x}{1-x-x^2} &= \frac{x}{(1-\alpha x)(1-\beta x)} = \frac{1}{\alpha-\beta}\left(\frac{1}{1-\alpha x} - \frac{1}{1-\beta x}\right)\\ &= \frac{1}{\alpha-\beta}\left(\sum_{n\geq 0}\alpha^n x^n - \sum_{n\geq 0}\beta^n x^n\right).\end{aligned}$$

By equating the coefficient of x^n, we have

$$F_n = \frac{\alpha^n - \beta^n}{\alpha - \beta},$$

which is exactly as desired. ■

In general, we may consider the sequence $\{G_n\}_{n\ge 0}$ with $G_0 = a$, $G_1 = b$, and for $n \ge 2$, $G_n = sG_{n-1} + tG_{n-2}$, where a, b, s and t are fixed.

Theorem 9.4 We have

$$\sum_{n\ge 0} G_n x^n = \frac{a+bx-asx}{1-sx-tx^2}. \tag{9.4}$$

Proof. Note that

$$\sum_{n\ge 2} G_n x^n = sx \sum_{n\ge 1} G_n x^n + tx^2 \sum_{n\ge 0} G_n x^n.$$

Thus,

$$\sum_{n\ge 0} G_n x^n - (a+bx) = sx \left(\sum_{n\ge 0} G_n x^n - a \right) + tx^2 \sum_{n\ge 0} G_n x^n,$$

yielding the desired result. ■

For example, the *Lucas numbers* L_n, which were introduced by the French mathematician François Lucas, are given by $L_0 = 2$, $L_1 = 1$, and for $n \ge 2$, $L_n = L_{n-1} + L_{n-2}$.

Theorem 9.5 We have

$$\sum_{n\ge 0} L_n x^n = \frac{2-x}{1-x-x^2}. \tag{9.5}$$

In particular, for $n \ge 0$,

$$L_n = \left(\frac{1+\sqrt{5}}{2}\right)^n + \left(\frac{1-\sqrt{5}}{2}\right)^n. \tag{9.6}$$

Proof. The first part is the $(a,b,s,t) = (2,1,1,1)$ case of Theorem 9.4. For the second part, we still write $\alpha = \frac{1+\sqrt{5}}{2}$ and $\beta = \frac{1-\sqrt{5}}{2}$. Then

$$\frac{2-x}{1-x-x^2} = \frac{1}{1-\alpha x} + \frac{1}{1-\beta x} = \sum_{n\ge 0} \alpha^n x^n + \sum_{n\ge 0} \beta^n x^n.$$

Equating the coefficient of x^n implies the desired result. ■

9.4 Compositions

Generating functions are of significant use in combinatorics. Here, we will take compositions as an example.

Definition 9.5 A *composition* of a positive integer n is a way of writing n as the sum of a sequence of positive integers, and the order of these summands matters.

■ **Example 9.3** There are four compositions of 3, namely, 3, $2+1$, $1+2$ and $1+1+1$. ■

Theorem 9.6 There are 2^{n-1} compositions of n.

Proof. We represent the integer n by n nodes in a row. Then there are $n-1$ gaps between consecutive nodes. Now, let us choose to place a stick at each gap or not, and there are 2^{n-1} choices. Each choice will induce a unique composition of n by counting the number of nodes between each pair of consecutive sticks while we assume that there are two invisible sticks at the two ends. Therefore, there are 2^{n-1} compositions of n.

• • | • • • | • • | • | •

For instance, the above diagram gives $2+3+2+1+1$, which is a composition of 9. ■

Is it possible to avoid such a combinatorial argument?

Theorem 9.7 Let $c(k,n)$ count the number of compositions of n into k parts. Then

$$\sum_{n\ge 1} c(k,n)x^n = \left(\frac{x}{1-x}\right)^k. \tag{9.7}$$

Proof. Let us consider the product

$$(x+x^2+\cdots)^k = (x+x^2+\cdots)(x+x^2+\cdots)\cdots(x+x^2+\cdots),$$

where there are k multiplicands. If we expand this product, then the terms are of the form $x^{n_1+n_2+\cdots+n_k} =: x^n$ where each x^{n_i} comes from the i-th multiplicand. Also, this term corresponds to a unique composition of n, given by $n = n_1+n_2+\cdots+n_k$, and there are exactly k parts in this composition. Hence,

$$\sum_{n\ge 1} c(k,n)x^n = (x+x^2+\cdots)^k = \left(\frac{x}{1-x}\right)^k,$$

as required. ■

Theorem 9.8 Let $c(n)$ count the number of compositions of n. Then

$$\sum_{n\ge 1} c(n)x^n = \frac{x}{1-2x}. \tag{9.8}$$

In particular, $c(n) = 2^{n-1}$.

Proof. For the first part, we deduce from Theorem 9.7 that

$$\sum_{n\ge 1} c(n)x^n = \sum_{k\ge 1}\sum_{n\ge 1} c(k,n)x^n = \sum_{k\ge 1}\left(\frac{x}{1-x}\right)^k = \frac{\frac{x}{1-x}}{1-\frac{x}{1-x}} = \frac{x}{1-2x}.$$

Further, $\frac{x}{1-2x} = \sum_{n\ge 1} 2^{n-1}x^n$. By equating the coefficient of x^n, we arrive at the second part. ■

10. Integer partitions

10.1 Integer partitions

Integer partitions can be viewed as a twin sibling of compositions.

Definition 10.1 An *integer partition* or a *partition* of a natural number n is a way of writing n as the sum of a sequence of positive integers, and the order of these summands does ***not*** matter. We usually denote by $p(n)$ the number of partitions of n, and call $p(n)$ the *partition function*.

R Since for a partition $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_\ell)$ of n, the order of these positive integers does not matter, we usually assume that they are in **weakly decreasing** order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_\ell$, as a representative. We also often write a partition as $\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_\ell$.

■ **Example 10.1** There are five partitions of 4, namely, 4, $3+1$, $2+2$, $2+1+1$ and $1+1+1+1$. Therefore, $p(4) = 5$. ■

Definition 10.2 Given a partition $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_\ell)$ of n, usually written as $\lambda \vdash n$, we call each λ_i a *part* of λ; call $n = \lambda_1 + \lambda_2 + \cdots + \lambda_\ell$ the *size* of λ, denoted by $|\lambda|$; and call the number ℓ of parts the *length* of λ, denoted by $\ell(\lambda)$.

R We assume that 0 has an empty partition, written as $\varnothing$, so $p(0) = 1$. For the empty partition $\varnothing$, we have $|\varnothing| = 0$ and $\ell(\varnothing) = 0$. Such a convention is beneficial when dealing with generating functions.

10.2 Generating function for partitions

Another convenient way to represent partitions is through the *frequency notation*. Given a partition λ, for each positive integer k, we may count the number f_k of occurrences of k among the parts in λ, and we call f_k the *frequency* of k. Hence, we may represent λ in the frequency notation $1^{f_1}2^{f_2}3^{f_3}\cdots$, and we often omit the integers whose frequency is zero.

■ **Example 10.2** The partition $6+6+5+3+3+3+2+1+1+1+1+1$ has the frequency notation $6^2 5^1 3^3 2^1 1^5$. ■

R When using the frequency notation, it is necessary to avoid confusion with products of powers.

Taking advantage of the frequency notation, it is easy to determine the generating function of $p(n)$.

Theorem 10.1 Let $p_{\leq N}(n)$ count the number of partitions of n with the largest part at most N. We have

$$\sum_{n\geq 0} p_{\leq N}(n)q^n = \prod_{k=1}^{N} \frac{1}{1-q^k}. \tag{10.1}$$

Proof. We expand the multiplicand

$$\frac{1}{1-q^k} = 1+q^k+q^{2k}+q^{3k}+\cdots = q^{0\cdot k}+q^{1\cdot k}+q^{2\cdot k}+q^{3\cdot k}+\cdots.$$

Hence, each term $q^{f_k\cdot k}$ enumerates the cases where the frequency of k is f_k for f_k a nonnegative integer. Further, if we expand the finite product $\prod_{k=1}^{N}\frac{1}{1-q^k}$, its terms are of the form $q^{f_1\cdot 1+f_2\cdot 2+\cdots+f_N\cdot N}$, corresponding to a unique partition with frequency notation $1^{f_1}2^{f_2}\cdots N^{f_N}$, which also restricts the largest part to be at most N. ■

Letting $N\to\infty$, we immediately see that the generating function of $p(n)$ is given by an infinite product.

Theorem 10.2 We have

$$\sum_{n\geq 0} p(n)q^n = \prod_{k\geq 1} \frac{1}{1-q^k}. \tag{10.2}$$

We may also apply some additional restrictions to the parts.

Theorem 10.3 For any positive integers $0<a\leq m$, let $p_{a,m}(n)$ count the number of partitions of n with parts congruent to a modulo m. We have

$$\sum_{n\geq 0} p_{a,m}(n)q^n = \prod_{k\geq 0} \frac{1}{1-q^{km+a}}. \tag{10.3}$$

Proof. Note that

$$\sum_{n\geq 0} p_{a,m}(n)q^n = \prod_{k\geq 0}\left(q^{0\cdot(km+a)}+q^{1\cdot(km+a)}+q^{2\cdot(km+a)}+\cdots\right) = \prod_{k\geq 0}\frac{1}{1-q^{km+a}},$$

as required. ■

Theorem 10.4 For any positive integer s, let $p_{[s]}(n)$ count the number of partitions of n in which each distinct part appears at most s times, i.e. the frequency satisfies $f_k\leq s$ for each k. We have

$$\sum_{n\geq 0} p_{[s]}(n)q^n = \prod_{k\geq 1} \frac{1-q^{(s+1)k}}{1-q^k}. \tag{10.4}$$

Proof. Note that

$$\sum_{n\geq 0} p_{[s]}(n)q^n = \prod_{k\geq 1}\left(q^{0\cdot k}+q^{1\cdot k}+\cdots+q^{s\cdot k}\right) = \prod_{k\geq 1}\frac{\left(1-q^k\right)\left(1+q+\cdots+q^{sk}\right)}{1-q^k} = \prod_{k\geq 1}\frac{1-q^{(s+1)k}}{1-q^k},$$

as required. ■

10.3 "Odd partitions" vs "Distinct partitions"

Definition 10.3 A partition is called an *odd partition* if all its parts are odd integers, and a partition is called an *even partition* if all its parts are even integers. We denote by $p_o(n)$ the number of odd partitions of n, and by $p_e(n)$ the number of even partitions of n.

Taking $m=2$, and then $a=1$ and 2, respectively, in Theorem 10.3, we have the following generating function identities.

Theorem 10.5 We have

$$\sum_{n\geq 0} p_o(n)q^n = \prod_{k\geq 1}\frac{1}{1-q^{2k-1}}, \tag{10.5}$$

$$\sum_{n\geq 0} p_e(n)q^n = \prod_{k\geq 1}\frac{1}{1-q^{2k}}. \tag{10.6}$$

Definition 10.4 A partition is called a *distinct partition* if all its parts are pairwise distinct. We denote by $p_D(n)$ the number of distinct partitions of n.

From the proof of Theorem 10.4 with $s=1$, the following generating function identity holds.

Theorem 10.6 We have

$$\sum_{n\geq 0} p_D(n)q^n = \prod_{k\geq 1}\left(1+q^k\right). \tag{10.7}$$

Euler established a well-known result on odd partitions and distinct partitions.

Theorem 10.7 (Euler). For $n\geq 0$, we have $p_o(n)=p_D(n)$.

Proof. It suffices to show that $p_o(n)$ and $p_D(n)$ have the same generating function:

$$\sum_{n\geq 0} p_o(n)q^n = \prod_{k\geq 1}\frac{1}{1-q^{2k-1}} = \prod_{k\geq 1}\frac{1}{1-q^{2k-1}}\frac{1-q^{2k}}{1-q^{2k}} = \prod_{k\geq 1}\frac{1-q^{2k}}{1-q^k} = \prod_{k\geq 1}\left(1+q^k\right) = \sum_{n\geq 0} p_D(n)q^n,$$

as required. ■

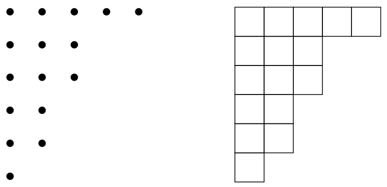
10.4 Ferrers diagrams

We may also represent partitions in a graphical way.

Definition 10.5 A *Ferrers diagram* represents a partition as patterns of dots, with the n-th row having the same number of dots as the n-th part of the partition. If we replace these dots with squares, the graph is often called a *Young diagram*.

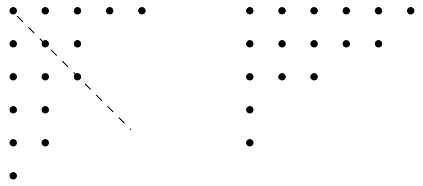
R Ferrers diagrams are named after the British mathematician Norman Macleod Ferrers, and Young diagrams are named after the British mathematician Alfred Young.

■ **Example 10.3** The graphical representations of the partition $5+3+3+2+2+1$ are given as follows — Ferrers diagram (left) and Young diagram (right): ■



Definition 10.6 Given a partition λ, its *conjugate partition*, denoted by λ^T, is the partition whose Ferrers diagram is obtained by flipping the diagram of λ along its main diagonal.

■ **Example 10.4** For the partition $\lambda = 5+3+3+2+2+1$, its conjugate is $\lambda^\mathsf{T} = 6+5+3+1+1$. ■



Theorem 10.8 Let $p(N,n)$ count the number of partitions of n with at most N parts. We have

$$\sum_{n\geq 0} p(N,n)q^n = \prod_{k=1}^{N} \frac{1}{1-q^k}. \tag{10.8}$$

Proof. Note that for any partition with at most N parts, its conjugate is a partition with the largest part at most N. Hence, $p(N,n) = p_{\leq N}(n)$. Recalling Theorem 10.1 gives the desired result. ■

10.5 Euler's summations

Note that the above generating functions are represented in the product form. Now we introduce the *q-Pochhammer symbols* for notational brevity.

Definition 10.7 Let $q \in \mathbb{C}$ be such that $|q| < 1$. Let $n \in \mathbb{N}$. The *q-Pochhammer symbols* are given by

$$(A;q)_n := \prod_{k=0}^{n-1}(1-Aq^k),$$
$$(A;q)_\infty := \prod_{k\geq 0}(1-Aq^k).$$

We first present refinements of Theorems 10.2 and 10.6.

Theorem 10.9 Let $\mathscr{P}$ be the set of partitions and $\mathscr{D}$ be the set of distinct partitions. We have

$$\sum_{\lambda\in\mathscr{P}} z^{\ell(\lambda)}q^{|\lambda|} = \frac{1}{(zq;q)_\infty}, \tag{10.9}$$
$$\sum_{\lambda\in\mathscr{D}} z^{\ell(\lambda)}q^{|\lambda|} = (-zq;q)_\infty. \tag{10.10}$$

Proof. We have

$$\sum_{\lambda\in\mathscr{P}} z^{\ell(\lambda)}q^{|\lambda|} = \prod_{k\geq 1}\left(1+zq^k+z^2q^{2k}+\cdots\right) = \prod_{k\geq 1}\frac{1}{1-zq^k} = \frac{1}{(zq;q)_\infty}.$$

Similarly,

$$\sum_{\lambda\in\mathscr{D}} z^{\ell(\lambda)}q^{|\lambda|} = \prod_{k\geq 1}\left(1+zq^k\right) = (-zq;q)_\infty,$$

as required. ■

Now, our objective is two important summation formulas due to Euler.

Theorem 10.10 (Euler's Summations). We have

$$\sum_{k\geq 0}\frac{z^kq^k}{(q;q)_k} = \frac{1}{(zq;q)_\infty}, \tag{10.11}$$

$$\sum_{k\geq 0}\frac{z^kq^{\frac{k(k+1)}{2}}}{(q;q)_k} = (-zq;q)_\infty. \tag{10.12}$$

Proof. For Euler's first summation, we consider partitions $\lambda = (\lambda_1,\lambda_2,\ldots,\lambda_k)\in\mathscr{P}$ with exactly k parts. Then $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_k\geq 1$. Now we construct a new partition $\lambda' = (\lambda_1',\lambda_2',\ldots,\lambda_k')$ with $\lambda_i' = \lambda_i - 1$. Noting that $\lambda_1'\geq\lambda_2'\geq\cdots\geq\lambda_k'\geq 0$, we find that λ' is a partition with at most k parts. Since $|\lambda| = |\lambda'|+k$, we have

$$\sum_{\lambda\in\mathscr{P}} z^{\ell(\lambda)}q^{|\lambda|} = \sum_{k\geq 0} z^kq^k\sum_{n\geq 0}p(k,n)q^n = \sum_{k\geq 0}\frac{z^kq^k}{(q;q)_k},$$

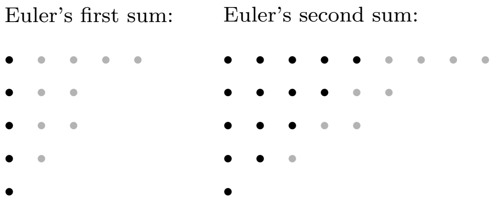
where we make use of Theorem 10.8. Recalling (10.9) gives what we want.

For Euler's second summation, we consider partitions $\pi = (\pi_1,\pi_2,\ldots,\pi_k)\in\mathscr{D}$ with exactly k parts. Then $\pi_1>\pi_2>\cdots>\pi_k\geq 1$. Now, we construct a new partition $\pi' = (\pi_1',\pi_2',\ldots,\pi_k')$ with $\pi_i' = \pi_i-(k+1-i)$. Noting that $\pi_1'\geq\pi_2'\geq\cdots\geq\pi_k'\geq 0$, we find that π' is a partition with at most k parts. Since $|\pi| = |\pi'|+(1+2+\cdots+k) = |\pi'|+\frac{k(k+1)}{2}$, we have

$$\sum_{\pi\in\mathscr{D}} z^{\ell(\pi)}q^{|\pi|} = \sum_{k\geq 0} z^kq^{\frac{k(k+1)}{2}}\sum_{n\geq 0}p(k,n)q^n = \sum_{k\geq 0}\frac{z^kq^{\frac{k(k+1)}{2}}}{(q;q)_k},$$

where we also use Theorem 10.8. Recalling (10.10) implies the desired result. ■

R The above proof can also be understood graphically.

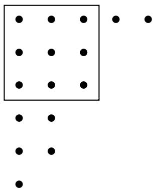


10.6 Durfee squares

From the Ferrers diagram of a partition, another important concept can be introduced.

Definition 10.8 Given a partition, its *Durfee square* is the largest square contained in its Ferrers diagram.

R Durfee squares are named after the American mathematician William Pitt Durfee, a student of James Joseph Sylvester.



■ **Example 10.5** The partition $5+3+3+2+2+1$ has a Durfee square of size 3, as shown in the Ferrers diagram. ■

Theorem 10.11 We have

$$\sum_{k\geq 0}\frac{q^{k^2}}{(q;q)_k^2}=\frac{1}{(q;q)_\infty}. \tag{10.13}$$

Proof. We consider partitions λ whose Durfee square is of size k. Note that below the Durfee square, we have a partition μ with the largest part at most k; and that to the right of the Durfee square, we have a partition ν with at most k parts. Since $|\lambda|=|\mu|+|\nu|+k^2$ where k^2 is contributed by the Durfee square, we have

$$\sum_{\lambda\in\mathscr{P}}q^{|\lambda|}=\sum_{k\geq 0}q^{k^2}\left(\sum_{n\geq 0}p_{\leq k}(n)q^n\right)\left(\sum_{n\geq 0}p(k,n)q^n\right)=\sum_{k\geq 0}\frac{q^{k^2}}{(q;q)_k^2},$$

where we consult Theorems 10.1 and 10.8. Finally, the desired identity follows from Theorem 10.2. ■

11. Basic q-series

11.1 q-Binomial series

We start with an identity due to the French mathematician Augustin-Louis Cauchy, which is also known as the *q-binomial series*.

Theorem 11.1 (q-Binomial Series). For $|q| < 1$ and $|t| < 1$,

$$\sum_{n\geq 0} \frac{(a;q)_n t^n}{(q;q)_n} = \frac{(at;q)_\infty}{(t;q)_\infty}. \tag{11.1}$$

R Taking $a = q^N$ in (11.1) with N a positive integer gives $\sum_{n\geq 0} \frac{(q^N;q)_n t^n}{(q;q)_n} = \frac{1}{(t;q)_N}$. Further letting $q \to 1^-$ implies that

$$\sum_{n\geq 0} \binom{N+n-1}{n} t^n = (1-t)^{-N}.$$

This provides an instance of the binomial theorem for negative powers.

Proof. Let us define

$$F(t) := \frac{(at;q)_\infty}{(t;q)_\infty}.$$

Note that as a function of t, $F(t)$ is analytic inside $|t| < 1$. Hence, we may expand $F(t)$ as a power series in t, i.e.

$$F(t) = \sum_{n\geq 0} f_n t^n.$$

Clearly, $f_0 = F(0) = 1$. Further, we have

$$F(tq) = \frac{(atq;q)_\infty}{(tq;q)_\infty} = \frac{1-t}{1-at} \cdot \frac{(at;q)_\infty}{(t;q)_\infty} = \frac{1-t}{1-at} \cdot F(t).$$

Since $F(tq) = \sum_{n\geq 0} f_n (tq)^n = \sum_{n\geq 0} \left(f_n q^n\right) t^n$, it follows that

$$(1-at) \sum_{n\geq 0} (f_n q^n) t^n = (1-t) \sum_{n\geq 0} f_n t^n.$$

Hence, for $n \geq 1$, we equate the coefficients of q^n on both sides of the above and obtain

$$f_n q^n - a f_{n-1} q^{n-1} = f_n - f_{n-1},$$

or

$$f_n = \frac{1 - aq^{n-1}}{1 - q^n} \cdot f_{n-1}.$$

Iterating the above gives

$$f_n = \frac{(a;q)_n}{(q;q)_n}$$

for $n \geq 0$. Substituting these coefficients back into $F(t) = \sum_{n\geq 0} f_n t^n$ confirms the required result. ■

R Euler's summations (10.11) and (10.12) are indeed special cases of the q-binomial series. For (10.11), we simply take $a = 0$ and $t = zq$ in (11.1), and note that for $n \in \mathbb{N} \cup \{\infty\}$, $(0;q)_n = (1-0)(1-0\cdot q)\cdots(1-0\cdot q^{n-1}) = 1$. For (10.12), we need the following trickier observation: for any nonnegative integer n,

$$\lim_{\tau\to 0} (a/\tau;q)_n \tau^n = \lim_{\tau\to 0} \tau^n \prod_{k=0}^{n-1} (1 - aq^k/\tau) = \lim_{\tau\to 0} \prod_{k=0}^{n-1} (\tau - aq^k) = \prod_{k=0}^{n-1} (-aq^k) = (-a)^n q^{\frac{n(n-1)}{2}}.$$

Now, in (11.1), we take $a \mapsto -zq/t$ and then let $t \to 0$. Therefore, (10.12) follows.

11.2 Heine's transformations

The q-binomial series serves as a key to many basic hypergeometric identities. Among these, *Heine's fundamental transformations* are of substantial significance.

Theorem 11.2 (Heine's Transformations). Let $|q| < 1$ and $|t| < 1$. For $|b| < 1$,

$$\sum_{n\geq 0} \frac{(a;q)_n (b;q)_n t^n}{(q;q)_n (c;q)_n} = \frac{(b;q)_\infty (at;q)_\infty}{(c;q)_\infty (t;q)_\infty} \sum_{n\geq 0} \frac{(c/b;q)_n (t;q)_n b^n}{(q;q)_n (at;q)_n}; \tag{11.2}$$

For $|c| < |b|$,

$$\sum_{n\geq 0} \frac{(a;q)_n (b;q)_n t^n}{(q;q)_n (c;q)_n} = \frac{(c/b;q)_\infty (bt;q)_\infty}{(c;q)_\infty (t;q)_\infty} \sum_{n\geq 0} \frac{(abt/c;q)_n (b;q)_n (c/b)^n}{(q;q)_n (bt;q)_n}; \tag{11.3}$$

For $|abt| < |c|$,

$$\sum_{n\geq 0} \frac{(a;q)_n (b;q)_n t^n}{(q;q)_n (c;q)_n} = \frac{(abt/c;q)_\infty}{(t;q)_\infty} \sum_{n\geq 0} \frac{(c/a;q)_n (c/b;q)_n (abt/c)^n}{(q;q)_n (c;q)_n}. \tag{11.4}$$

R These transformation formulas were first studied by the German mathematician Eduard Heine (*J. Reine Angew. Math.* **32** (1846), 210–212).

Proof. We begin with a trivial observation that for any nonnegative integer n,

$$\frac{(\alpha;q)_n}{(\beta;q)_n} = \frac{(\alpha;q)_\infty}{(\beta;q)_\infty} \frac{(\beta q^n;q)_\infty}{(\alpha q^n;q)_\infty}.$$

Now, for (11.2), we have

$$\begin{aligned}\sum_{n\ge0}\frac{(a;q)_n(b;q)_nt^n}{(q;q)_n(c;q)_n}&=\frac{(b;q)_\infty}{(c;q)_\infty}\sum_{n\ge0}\frac{(a;q)_nt^n}{(q;q)_n}\cdot\frac{(cq^n;q)_\infty}{(bq^n;q)_\infty}\\ \text{(by (11.1))}\ &=\frac{(b;q)_\infty}{(c;q)_\infty}\sum_{n\ge0}\frac{(a;q)_nt^n}{(q;q)_n}\sum_{m\ge0}\frac{(c/b;q)_m(bq^n)^m}{(q;q)_m}\\ &=\frac{(b;q)_\infty}{(c;q)_\infty}\sum_{m\ge0}\frac{(c/b;q)_mb^m}{(q;q)_m}\sum_{n\ge0}\frac{(a;q)_n(tq^m)^n}{(q;q)_n}\\ \text{(by (11.1))}\ &=\frac{(b;q)_\infty}{(c;q)_\infty}\sum_{m\ge0}\frac{(c/b;q)_mb^m}{(q;q)_m}\cdot\frac{(atq^m;q)_\infty}{(tq^m;q)_\infty}\\ &=\frac{(b;q)_\infty}{(c;q)_\infty}\frac{(at;q)_\infty}{(t;q)_\infty}\sum_{m\ge0}\frac{(c/b;q)_mb^m}{(q;q)_m}\cdot\frac{(t;q)_m}{(at;q)_m},\end{aligned}$$

as required. For (11.3), we first take $(a,b,c,t)\mapsto(t,c/b,at,b)$ in (11.2). Then

$$\sum_{n\ge0}\frac{(t;q)_n(c/b;q)_nb^n}{(q;q)_n(at;q)_n}=\frac{(c/b;q)_\infty(bt;q)_\infty}{(at;q)_\infty(b;q)_\infty}\sum_{n\ge0}\frac{(abt/c;q)_n(b;q)_n(c/b)^n}{(q;q)_n(bt;q)_n}.$$

Substituting the above into the right-hand side of (11.2) gives

$$\sum_{n\ge0}\frac{(a;q)_n(b;q)_nt^n}{(q;q)_n(c;q)_n}=\frac{(b;q)_\infty(at;q)_\infty}{(c;q)_\infty(t;q)_\infty}\cdot\frac{(c/b;q)_\infty(bt;q)_\infty}{(at;q)_\infty(b;q)_\infty}\sum_{n\ge0}\frac{(abt/c;q)_n(b;q)_n(c/b)^n}{(q;q)_n(bt;q)_n},$$

which is exactly (11.3). Finally, for (11.4), we take $(a,b,c,t)\mapsto(b,abt/c,bt,c/b)$ in (11.2). Then

$$\sum_{n\ge0}\frac{(b;q)_n(abt/c;q)_n(c/b)^n}{(q;q)_n(bt;q)_n}=\frac{(abt/c;q)_\infty(c;q)_\infty}{(bt;q)_\infty(c/b;q)_\infty}\sum_{n\ge0}\frac{(c/a;q)_n(c/b;q)_n(abt/c)^n}{(q;q)_n(c;q)_n}.$$

Substituting the above into the right-hand side of (11.3) gives

$$\sum_{n\ge0}\frac{(a;q)_n(b;q)_nt^n}{(q;q)_n(c;q)_n}=\frac{(c/b;q)_\infty(bt;q)_\infty}{(c;q)_\infty(t;q)_\infty}\cdot\frac{(abt/c;q)_\infty(c;q)_\infty}{(bt;q)_\infty(c/b;q)_\infty}\sum_{n\ge0}\frac{(c/a;q)_n(c/b;q)_n(abt/c)^n}{(q;q)_n(c;q)_n},$$

thereby confirming (11.4). ■

As an important consequence of Heine's transformations, we have the *q-Gauss summation.*

Corollary 11.3 (q-Gauss Summation). For $|q|<1$ and $|c|<|ab|$,

$$\sum_{n\ge0}\frac{(a;q)_n(b;q)_n}{(q;q)_n(c;q)_n}\left(\frac{c}{ab}\right)^n=\frac{(c/a;q)_\infty(c/b;q)_\infty}{(c;q)_\infty(c/(ab);q)_\infty}.\tag{11.5}$$

Proof. In Heine's first transformation (11.2), we take $t\mapsto c/(ab)$. Then

$$\begin{aligned}\sum_{n\ge0}\frac{(a;q)_n(b;q)_n}{(q;q)_n(c;q)_n}\left(\frac{c}{ab}\right)^n&=\frac{(b;q)_\infty(c/b;q)_\infty}{(c;q)_\infty(c/(ab);q)_\infty}\sum_{n\ge0}\frac{(c/b;q)_n(c/(ab);q)_nb^n}{(q;q)_n(c/b;q)_n}\\ &=\frac{(b;q)_\infty(c/b;q)_\infty}{(c;q)_\infty(c/(ab);q)_\infty}\sum_{n\ge0}\frac{(c/(ab);q)_nb^n}{(q;q)_n}\\ \text{(by (11.1))}\ &=\frac{(b;q)_\infty(c/b;q)_\infty}{(c;q)_\infty(c/(ab);q)_\infty}\cdot\frac{(c/a;q)_\infty}{(b;q)_\infty},\end{aligned}$$

which leads to the required identity. ■

R It is worth pointing out that (10.13) is a special case of the q-Gauss summation by first taking $(a,b,c)\mapsto(1/\tau,1/\tau,q)$ in (11.5) and then letting $\tau\to 0$.

11.3 Jacobi's triple product identity

Let us take $z\mapsto z/q$ in Euler's two summation formulas (10.11) and (10.12):

$$\sum_{k\ge 0}\frac{z^k}{(q;q)_k}=\frac{1}{(z;q)_\infty}, \tag{11.6}$$

$$\sum_{k\ge 0}\frac{z^k q^{\frac{k(k-1)}{2}}}{(q;q)_k}=(-z;q)_\infty. \tag{11.7}$$

From the discussions in the final remark in §11.1, we see that under the assumption of $|q|<1$, (11.6) is true for $|z|<1$ and (11.7) is true for any complex z.

Now we shall use them to prove one of the most important q-series identities — *Jacobi's triple product identity*, named after the German mathematician Carl Gustav Jacob Jacobi.

Theorem 11.4 (Jacobi's Triple Product Identity). For $|q|<1$ and $z\neq 0$,

$$\sum_{n=-\infty}^{\infty}(-z)^n q^{\frac{n(n-1)}{2}}=(z;q)_\infty(q/z;q)_\infty(q;q)_\infty. \tag{11.8}$$

Proof. We start with (11.7) and deduce that

$$(-z;q)_\infty=\sum_{k\ge 0}\frac{z^k q^{\frac{k(k-1)}{2}}}{(q;q)_k}=\frac{1}{(q;q)_\infty}\sum_{k\ge 0}z^k q^{\frac{k(k-1)}{2}}(q^{k+1};q)_\infty.$$

Note that for j a nonpositive integer, in $(q^j;q)_\infty=(1-q^j)(1-q^{j+1})\cdots$, one of the factors is $(1-q^0)=(1-1)=0$, thereby yielding $(q^j;q)_\infty=0$ for every such j. It turns out that the above summation can be extended as a bilateral one,

$$\begin{aligned}
(-z;q)_\infty &= \frac{1}{(q;q)_\infty}\sum_{k=-\infty}^{\infty}z^k q^{\frac{k(k-1)}{2}}(q^{k+1};q)_\infty\\
\text{(by (11.7))} &= \frac{1}{(q;q)_\infty}\sum_{k=-\infty}^{\infty}z^k q^{\frac{k(k-1)}{2}}\sum_{\ell\ge 0}\frac{(-q^{k+1})^\ell q^{\frac{\ell(\ell-1)}{2}}}{(q;q)_\ell}\\
&= \frac{1}{(q;q)_\infty}\sum_{k=-\infty}^{\infty}\sum_{\ell\ge 0}\frac{(-1)^\ell\cdot z^k\cdot q^{\frac{\ell(\ell-1)}{2}+\frac{k(k-1)}{2}+(k+1)\ell}}{(q;q)_\ell}\\
&= \frac{1}{(q;q)_\infty}\sum_{k=-\infty}^{\infty}\sum_{\ell\ge 0}\frac{(-1)^\ell\cdot z^k\cdot q^{\frac{(\ell+k)(\ell+k-1)}{2}+\ell}}{(q;q)_\ell}\\
&= \frac{1}{(q;q)_\infty}\sum_{\ell\ge 0}\frac{(-1)^\ell z^{-\ell}q^\ell}{(q;q)_\ell}\sum_{k=-\infty}^{\infty}z^{\ell+k}q^{\frac{(\ell+k)(\ell+k-1)}{2}}\\
\text{(with } n=\ell+k) &= \frac{1}{(q;q)_\infty}\sum_{\ell\ge 0}\frac{(-1)^\ell z^{-\ell}q^\ell}{(q;q)_\ell}\sum_{n=-\infty}^{\infty}z^n q^{\frac{n(n-1)}{2}}\\
\text{(by (11.6))} &= \frac{1}{(q;q)_\infty}\frac{1}{(-q/z;q)_\infty}\sum_{n=-\infty}^{\infty}z^n q^{\frac{n(n-1)}{2}}.
\end{aligned}$$

Note that in the last equality, we should require $|q/z| < 1$, or $|z| > |q|$ to pertain the absolute convergence. However, the entire argument can be carried out again with z replaced by q/z. Namely, for $0 < |z| < 1$,

$$(-q/z;q)_\infty = \frac{1}{(q;q)_\infty(-z;q)_\infty}\sum_{n=-\infty}^{\infty} z^{-n}q^{\frac{n(n+1)}{2}} = \frac{1}{(q;q)_\infty(-z;q)_\infty}\sum_{n=-\infty}^{\infty} z^{n}q^{\frac{n(n-1)}{2}}.$$

Further, $\{z : |z| > |q|\} \cup \{z : 0 < |z| < 1\} = \mathbb{C}\backslash\{0\}$ since $|q| < 1$. We remark that a simpler way to get rid of the requirement that $|z| > |q|$ is by invoking analytic continuation. Finally, we derive from the above that for $z \neq 0$,

$$\sum_{n=-\infty}^{\infty} z^n q^{\frac{n(n-1)}{2}} = (-z;q)_\infty(-q/z;q)_\infty(q;q)_\infty,$$

thereby yielding the desired result by setting $z \mapsto -z$. ■

A direct consequence of Jacobi's triple product identity is *Euler's pentagonal number theorem.*

Corollary 11.5 (Euler's Pentagonal Number Theorem). For $|q| < 1$,

$$\sum_{n=-\infty}^{\infty} (-1)^n q^{\frac{n(3n-1)}{2}} = (q;q)_\infty. \tag{11.9}$$

Proof. In (11.8), we take $(z,q) \mapsto (q,q^3)$. Noting that $(q;q^3)_\infty(q^2;q^3)_\infty(q^3;q^3)_\infty = (q;q)_\infty$, we arrive at the desired result. ■

11.4 Ramanujan's theta function

An important object in the theory of q-series is the *theta function* introduced by the Indian mathematician Srinivasa Ramanujan.

Definition 11.1 *Ramanujan's general theta function* is defined as

$$f(a,b) := \sum_{n=-\infty}^{\infty} a^{\frac{n(n+1)}{2}} b^{\frac{n(n-1)}{2}} \qquad (|ab| < 1). \tag{11.10}$$

Theorem 11.6 For $|ab| < 1$,

$$f(a,b) = (-a;ab)_\infty(-b;ab)_\infty(ab;ab)_\infty. \tag{11.11}$$

Proof. This is (11.8) with $(z,q) \mapsto (-a,ab)$. ■

Two special cases of the general theta function are of particular interest.

Definition 11.2 *Ramanujan's classical theta functions* are defined as

$$\phi(q) := \sum_{n=-\infty}^{\infty} q^{n^2}, \tag{11.12}$$

$$\psi(q) := \sum_{n\geq 0} q^{\frac{n(n+1)}{2}}. \tag{11.13}$$

Theorem 11.7 We have

$$\phi(q) = \frac{(q^2;q^2)_\infty^5}{(q;q)_\infty^2(q^4;q^4)_\infty^2}, \tag{11.14}$$

$$\psi(q) = \frac{(q^2;q^2)_\infty^2}{(q;q)_\infty}, \tag{11.15}$$

$$\phi(-q) = \frac{(q;q)_\infty^2}{(q^2;q^2)_\infty}, \tag{11.16}$$

$$\psi(-q) = \frac{(q;q)_\infty(q^4;q^4)_\infty}{(q^2;q^2)_\infty}. \tag{11.17}$$

Proof. For (11.14), we note that

$$\phi(q) = \sum_{n=-\infty}^{\infty} q^{n^2} = f(q,q). \tag{11.18}$$

Hence, it follows from (11.11) that

$$\phi(q) = (-q;q^2)_\infty^2(q^2;q^2)_\infty = \frac{(q^2;q^4)_\infty^2}{(q;q^2)_\infty^2}(q^2;q^2)_\infty = \frac{(q^2;q^2)_\infty^2}{(q^4;q^4)_\infty^2}\frac{(q^2;q^2)_\infty^2}{(q;q)_\infty^2}(q^2;q^2)_\infty,$$

as required. For (11.15), we first show that

$$\sum_{n\ge 0} q^{\frac{n(n+1)}{2}} = \sum_{n=-\infty}^{\infty} q^{2n^2-n} = f(q,q^3). \tag{11.19}$$

To see this, for the left-hand side, we distinguish the parity of n and write n as $2k$ and $2k+1$ with $k \ge 0$. On the other hand, for the right-hand side, we separate n into $-k$ and $k+1$, also with $k \ge 0$. Then

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

and

$$\sum_{n=-\infty}^{\infty} q^{2n^2+n} = \sum_{k\ge 0} q^{2(-k)^2-(-k)} + \sum_{k\ge 0} q^{2(k+1)^2-(k+1)} = \sum_{k\ge 0} q^{k(2k+1)} + \sum_{k\ge 0} q^{(k+1)(2k+1)},$$

and thus they are equal. By (11.11), we have

$$\begin{aligned}\psi(q) &= (-q;q^4)_\infty(-q^3;q^4)_\infty(q^4;q^4)_\infty = (-q;q^2)_\infty(q^4;q^4)_\infty = \frac{(q^2;q^4)_\infty}{(q;q^2)_\infty}(q^4;q^4)_\infty \\ &= \frac{(q^2;q^2)_\infty}{(q^4;q^4)_\infty}\frac{(q^2;q^2)_\infty}{(q;q)_\infty}(q^4;q^4)_\infty,\end{aligned}$$

as required. Finally, for (11.16) and (11.17), we note that

$$(-q;-q)_\infty = (1+q)(1-q^2)(1+q^3)(1-q^4) = (-q;q^2)_\infty(q^2;q^2)_\infty.$$

Hence,

$$(-q;-q)_\infty = \frac{(q^2;q^2)_\infty^3}{(q;q)_\infty(q^4;q^4)_\infty}. \tag{11.20}$$

Taking $q \mapsto -q$ in (11.14) and (11.15), and making use of the above relation, the desired results follow. ■

12. Sums of squares (II)

12.1 Jacobi's identity

Here, we record another important implication of Jacobi's triple product identity.

Theorem 12.1 (Jacobi's Identity). For $|q|<1$,

$$\sum_{n\geq 0}(-1)^n(2n+1)q^{\frac{n(n+1)}{2}}=(q;q)_\infty^3. \tag{12.1}$$

Proof. Recall (11.8):

$$(z;q)_\infty(z^{-1}q;q)_\infty(q;q)_\infty=\sum_{n=-\infty}^{\infty}(-z)^n q^{\frac{n(n-1)}{2}}.$$

Note that the product side can be rewritten as

$$(z;q)_\infty(z^{-1}q;q)_\infty(q;q)_\infty=-(z-1)(zq;q)_\infty(z^{-1}q;q)_\infty(q;q)_\infty.$$

For the summation side, we distinguish n as $-k$ and $k+1$ with $k\geq 0$:

$$\begin{aligned}\sum_{n=-\infty}^{\infty}(-z)^n q^{\frac{n(n-1)}{2}}&=\sum_{k\geq 0}(-1)^k z^{-k}q^{\frac{k(k+1)}{2}}-\sum_{k\geq 0}(-1)^k z^{k+1}q^{\frac{k(k+1)}{2}}\\&=-\sum_{k\geq 0}(-1)^k\left(z^{k+1}-z^{-k}\right)q^{\frac{k(k+1)}{2}}.\end{aligned}$$

Hence,

$$(z-1)(zq;q)_\infty(z^{-1}q;q)_\infty(q;q)_\infty=\sum_{k\geq 0}(-1)^k\left(z^{k+1}-z^{-k}\right)q^{\frac{k(k+1)}{2}}.$$

Now, note that $z^{k+1}-z^{-k}=(z-1)(z^k+z^{k-1}+\cdots+z^{-k})$. We then divide by $z-1$ on both sides of the above and obtain

$$(zq;q)_\infty(z^{-1}q;q)_\infty(q;q)_\infty=\sum_{k\geq 0}(-1)^k\left(z^k+z^{k-1}+\cdots+z^{-k}\right)q^{\frac{k(k+1)}{2}}.$$

Finally, taking $z=1$ gives the desired result. ■

12.2 Lambert series

Definition 12.1 Let k be a fixed positive integer. For every natural number n, we denote by $r_k(n)$ the number of representations of n as $m_1^2+m_2^2+\cdots+m_k^2$ with all m_i integers, where representations differing only in the sign or order of the m_i shall be reckoned as distinct.

Example 12.1 We can represent 5 as

$$\begin{array}{cccc} 1^2+2^2, & (-1)^2+2^2, & 1^2+(-2)^2, & (-1)^2+(-2)^2, \\ 2^2+1^2, & 2^2+(-1)^2, & (-2)^2+1^2, & (-2)^2+(-1)^2. \end{array}$$

Hence, $r_2(5)=8$. ■

Theorem 12.2 Let $\phi(q)$ be Ramanujan's theta function as in (11.12). We have

$$1+\sum_{n\ge 1} r_k(n)q^n = \phi(q)^k. \tag{12.2}$$

Proof. This is a direct consequence of $\phi(q)=\sum_{n=-\infty}^{\infty} q^{n^2}$. ■

Now our object is to derive explicit formulas for $r_2(n)$ and $r_4(n)$. For this purpose, we require the knowledge of *Lambert series*, named after the Swiss–German mathematician Johann Heinrich Lambert.

Definition 12.2 A *Lambert series* is of the form

$$\sum_{k\ge 1}\frac{a_k q^k}{1-q^k},$$

where $\{a_k\}_{k\ge 1}$ is a sequence of complex numbers.

Theorem 12.3 Let

$$\sum_{n\ge 1} u_n q^n = \sum_{\substack{k\ge 1\\ k\equiv r \bmod m}} \frac{a_k q^k}{1-q^k}.$$

Then

$$u_n = \sum_{\substack{d\mid n\\ d\equiv r \bmod m}} a_d. \tag{12.3}$$

Proof. We expand the summand

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

Note that q^n appears in this series if and only if $k\mid n$. Since we are summing over all positive integers k with $k\equiv r \pmod m$ in the Lambert series, then to compute the coefficient u_n, we need to take into account all positive divisors d of n with $d\equiv r \pmod m$, and thus arrive at the required expression. ■

Lemma 12.4 We have

$$\frac{d}{dx}\prod_k f_k(x)=\left(\prod_k f_k(x)\right)\cdot\sum_k\left(\frac{1}{f_k(x)}\cdot\frac{d}{dx}f_k(x)\right). \tag{12.4}$$

Proof. Let $F(x)=\prod_k f_k(x)$. Note that $\frac{d}{dx}\log F(x)=\frac{F'(x)}{F(x)}$, where $F'(x)$ denotes the derivative of $F(x)$. Hence,

$$F'(x)=F(x)\cdot\frac{d}{dx}\log F(x)=F(x)\cdot\frac{d}{dx}\sum_k\log f_k(x)=F(x)\cdot\sum_k\frac{d}{dx}\log f_k(x)=F(x)\cdot\sum_k\frac{f_k'(x)}{f_k(x)},$$

as required. ■

R This lemma allows us to connect q-Pochhammer symbols with the Lambert series through differentiation. For instance,

$$q\cdot\frac{d}{dq}(q;q)_\infty=q\cdot\frac{d}{dq}\prod_{k\geq1}(1-q^k)=q(q;q)_\infty\sum_{k\geq1}\frac{-kq^{k-1}}{1-q^k}=-(q;q)_\infty\sum_{k\geq1}\frac{kq^k}{1-q^k}.$$

12.3 Jacobi's two-square formula

Theorem 12.5 (Jacobi's Two-Square Formula). For $n\geq1$,

$$r_2(n)=4\left(\sum_{\substack{d\mid n\\ d\equiv1 \bmod 4}}1-\sum_{\substack{d\mid n\\ d\equiv3 \bmod 4}}1\right). \tag{12.5}$$

Proof. We begin with Jacobi's identity (12.1):

$$\begin{aligned}(q;q)_\infty^3&=\sum_{n\geq0}(-1)^n(2n+1)q^{\frac{n(n+1)}{2}}=\sum_{k\geq0}(4k+1)q^{\frac{(2k)(2k+1)}{2}}-\sum_{k\geq0}(4k+3)q^{\frac{(2k+1)(2k+2)}{2}}\\&=\sum_{k\geq0}(4k+1)q^{2k^2+k}+\sum_{k\geq0}\big(4(-k-1)+1\big)q^{2(-k-1)^2+(-k-1)}\\&=\sum_{n=-\infty}^{\infty}(4n+1)q^{2n^2+n}.\end{aligned}$$

Hence,

$$\begin{aligned}(q;q)_\infty^3&=\left[\frac{d}{dz}\left(\sum_{n=-\infty}^{\infty}z^{4n+1}q^{2n^2+n}\right)\right]_{z=1}\\ \text{(by (11.8))}&=\left[\frac{d}{dz}\left(z(-z^{-4}q;q^4)_\infty(-z^4q^3;q^4)_\infty(q^4;q^4)_\infty\right)\right]_{z=1}\\ \text{(by (12.4))}&=(-q;q^4)_\infty(-q^3;q^4)_\infty(q^4;q^4)_\infty\left(1-\sum_{\substack{k\geq1\\ k\equiv1 \bmod 4}}\frac{4q^k}{1+q^k}+\sum_{\substack{k\geq1\\ k\equiv3 \bmod 4}}\frac{4q^k}{1+q^k}\right).\end{aligned}$$

In the proof of Theorem 11.7, we have shown that

$$\psi(q)=(-q;q^4)_\infty(-q^3;q^4)_\infty(q^4;q^4)_\infty.$$

Recalling (11.15) and (11.16), we have

$$\phi(-q)^2 = 1 - \sum_{\substack{k\geq 1\\ k\equiv 1 \bmod 4}} \frac{4q^k}{1+q^k} + \sum_{\substack{k\geq 1\\ k\equiv 3 \bmod 4}} \frac{4q^k}{1+q^k}.$$

Now we take $q \mapsto -q$ and derive that

$$\phi(q)^2 = 1 + \sum_{\substack{k\geq 1\\ k\equiv 1 \bmod 4}} \frac{4q^k}{1-q^k} - \sum_{\substack{k\geq 1\\ k\equiv 3 \bmod 4}} \frac{4q^k}{1-q^k}.$$

Finally, the required result follows by using (12.2) and (12.3). ■

R This proof comes from an unpublished work of the Australian mathematician Michael Hirschhorn. See also Hirschhorn's monograph *The power of q*, Sect. 2.3.

12.4 Jacobi's four-square formula

Theorem 12.6 (Jacobi's Four-Square Formula). For $n \geq 1$,

$$r_4(n) = 8 \sum_{\substack{d\mid n\\ d\not\equiv 0 \bmod 4}} d. \tag{12.6}$$

For its proof, we need a reformulation of $(q;q)_\infty^6$.

Lemma 12.7 We have

$$(q;q)_\infty^6 = \frac{1}{2}\sum_{s=-\infty}^{\infty} q^{s^2} \sum_{r=-\infty}^{\infty} (2r+1)^2 q^{r^2+r} - \frac{1}{2}\sum_{r=-\infty}^{\infty} q^{r^2+r} \sum_{s=-\infty}^{\infty} (2s)^2 q^{s^2}. \tag{12.7}$$

Proof. We note from Jacobi's identity (12.1) that

$$\begin{aligned}
&\sum_{n=-\infty}^{\infty} (-1)^n (2n+1) q^{\frac{n(n+1)}{2}}\\
&\qquad = \sum_{n\geq 0} (-1)^n (2n+1) q^{\frac{n(n+1)}{2}} + \sum_{n\geq 0} (-1)^{-n-1}\big(2(-n-1)+1\big) q^{\frac{(-n-1)((-n-1)+1)}{2}}\\
&\qquad = 2\sum_{n\geq 0} (-1)^n (2n+1) q^{\frac{n(n+1)}{2}}\\
&\qquad = 2(q;q)_\infty^3.
\end{aligned}$$

Hence,

$$(q;q)_\infty^6 = \frac{1}{4}\sum_{m,n=-\infty}^{\infty} (-1)^{m+n} (2m+1)(2n+1) q^{\frac{m(m+1)}{2}+\frac{n(n+1)}{2}}.$$

We may further split the sum into two parts, according to whether m and n have the same parity or not, and obtain

$$(q;q)_\infty^6 = \frac{1}{4}\sum_{\substack{m,n=-\infty\\ m\equiv n \bmod 2}}^{\infty} (2m+1)(2n+1) q^{\frac{m(m+1)}{2}+\frac{n(n+1)}{2}}$$

$$-\frac{1}{4}\sum_{\substack{m,n=-\infty\\ m\not\equiv n \bmod 2}}^{\infty}(2m+1)(2n+1)q^{\frac{m(m+1)}{2}+\frac{n(n+1)}{2}}.$$

For the first sum,

$$\sum_{\substack{m,n=-\infty\\ m\equiv n \bmod 2}}^{\infty}(2m+1)(2n+1)q^{\frac{m(m+1)}{2}+\frac{n(n+1)}{2}},$$

we make the following change of variables (so that r and s run over all integers):

$$\begin{cases} r=\frac{m+n}{2}\\ s=\frac{m-n}{2}\end{cases} \iff \begin{cases} m=r+s\\ n=r-s\end{cases}.$$

Similarly, for the second sum,

$$\sum_{\substack{m,n=-\infty\\ m\not\equiv n \bmod 2}}^{\infty}(2m+1)(2n+1)q^{\frac{m(m+1)}{2}+\frac{n(n+1)}{2}},$$

we make another change of variables:

$$\begin{cases} r=\frac{m-n-1}{2}\\ s=\frac{m+n+1}{2}\end{cases} \iff \begin{cases} m=r+s\\ n=s-r-1\end{cases}.$$

Thus,

$$\begin{aligned}(q;q)_\infty^6 &= \frac{1}{4}\sum_{r,s=-\infty}^{\infty}\big((2r+1)^2-(2s)^2\big)q^{r^2+r+s^2}-\frac{1}{4}\sum_{r,s=-\infty}^{\infty}\big((2s)^2-(2r+1)^2\big)q^{r^2+r+s^2}\\ &=\frac{1}{2}\sum_{r,s=-\infty}^{\infty}\big((2r+1)^2-(2s)^2\big)q^{r^2+r+s^2},\end{aligned}$$

as required. ■

Now we are in a position to prove Theorem 12.6.

Proof of Theorem 12.6. We first reformulate (12.7) and get

$$\begin{aligned}(q;q)_\infty^6 &= \frac{1}{2}\sum_{s=-\infty}^{\infty}q^{s^2}\sum_{r=-\infty}^{\infty}(2r+1)^2q^{r^2+r}-\frac{1}{2}\sum_{r=-\infty}^{\infty}q^{r^2+r}\sum_{s=-\infty}^{\infty}(2s)^2q^{s^2}\\ &=\frac{1}{2}\sum_{s=-\infty}^{\infty}q^{s^2}\sum_{r=-\infty}^{\infty}q^{r^2+r}+2\sum_{s=-\infty}^{\infty}q^{s^2}\sum_{r=-\infty}^{\infty}(r^2+r)q^{r^2+r}-2\sum_{r=-\infty}^{\infty}q^{r^2+r}\sum_{s=-\infty}^{\infty}s^2q^{s^2}\\ &=\frac{1}{2}\sum_{s=-\infty}^{\infty}q^{s^2}\sum_{r=-\infty}^{\infty}q^{r^2+r}+2\sum_{s=-\infty}^{\infty}q^{s^2}\cdot q\frac{d}{dq}\sum_{r=-\infty}^{\infty}q^{r^2+r}-2\sum_{r=-\infty}^{\infty}q^{r^2+r}\cdot q\frac{d}{dq}\sum_{s=-\infty}^{\infty}q^{s^2}.\end{aligned}$$

Since

$$\sum_{n=-\infty}^{\infty}q^{n^2}=\phi(q)$$

and

$$\sum_{n=-\infty}^{\infty}q^{n^2+n}=2\sum_{n\geq 0}q^{n^2+n}=2\psi(q^2),$$

we further have

$$(q;q)_\infty^6=\phi(q)\psi(q^2)+4\phi(q)\cdot q\frac{d}{dq}\psi(q^2)-4\psi(q^2)\cdot q\frac{d}{dq}\phi(q).$$

Now, by Jacobi's triple product identity (11.8),

$$\phi(q) = (-q;q^2)_\infty^2 (q^2;q^2)_\infty.$$

Also, by (11.15),

$$\psi(q^2) = \frac{(q^4;q^4)_\infty^2}{(q^2;q^2)_\infty} = \frac{(q^4;q^4)_\infty^2}{(q^2;q^2)_\infty} \frac{(q^2;q^4)_\infty^2}{(q^2;q^4)_\infty^2} = \frac{(q^2;q^2)_\infty}{(q^2;q^4)_\infty^2}.$$

It is a routine exercise by applying (12.4) to the above two relations that

$$q\frac{d}{dq}\phi(q) = \phi(q) \sum_{k\ge 1} \left(\frac{2(2k-1)q^{2k-1}}{1+q^{2k-1}} - \frac{2kq^{2k}}{1-q^{2k}} \right)$$

and

$$q\frac{d}{dq}\psi(q^2) = \psi(q^2) \sum_{k\ge 1} \left(\frac{2(4k-2)q^{4k-2}}{1-q^{4k-2}} - \frac{2kq^{2k}}{1-q^{2k}} \right).$$

Therefore,

$$(q;q)_\infty^6 = \phi(q)\psi(q^2) \left(1 + 8 \sum_{k\ge 1} \left(\frac{(4k-2)q^{4k-2}}{1-q^{4k-2}} - \frac{(2k-1)q^{2k-1}}{1+q^{2k-1}} \right) \right).$$

Recalling (11.14), (11.15) and (11.16), we have

$$\phi(-q)^4 = 1 + 8 \sum_{k\ge 1} \left(\frac{(4k-2)q^{4k-2}}{1-q^{4k-2}} - \frac{(2k-1)q^{2k-1}}{1+q^{2k-1}} \right).$$

Finally, we take $q \mapsto -q$ and derive that

$$\begin{aligned} \phi(q)^4 &= 1 + 8 \sum_{k\ge 1} \left(\frac{(4k-2)q^{4k-2}}{1-q^{4k-2}} + \frac{(2k-1)q^{2k-1}}{1-q^{2k-1}} \right) \\ &= 1 + 8 \sum_{\substack{k\ge 1 \\ k\not\equiv 0 \text{ mod } 4}} \frac{kq^k}{1-q^k}. \end{aligned}$$

The desired result follows by applying (12.2) and (12.3). ■

This proof is also attributed to Hirschhorn (*Proc. Amer. Math. Soc.* **101** (1987), no. 3, 436–438).

13. Arithmetic functions

13.1 Arithmetic functions

In the previous lectures, we have witnessed functions like the "sum-of-squares" functions $r_k(n)$ that are defined on the positive integers. Such functions are of particular interest in the study of number theory.

Definition 13.1 An *arithmetic function* is a complex-valued function that is defined on the positive integers.

R In G. H. Hardy and E. M. Wright's *Introduction*, they also include in their definition the requirement that an arithmetical function "expresses some arithmetical property of [each positive integer]."

Recall that we have also encountered multiplicative functions such as Euler's totient function and completely multiplicative functions such as the Legendre symbol restricted to the positive integers.

Definition 13.2 An arithmetic function f is

(i) *multiplicative* if $f(1) = 1$ and $f(mn) = f(m)f(n)$ for all positive integers m and n with $(m,n) = 1$;

(ii) *completely multiplicative* if $f(1) = 1$ and $f(mn) = f(m)f(n)$ for all positive integers m and n.

R Observe that for any multiplicative function f, we have $f(1) = f(1\cdot 1) = f(1)\cdot f(1)$. Hence, there are only two possibilities of $f(1)$, namely, $f(1) = 1$ or $f(1) = 0$. However, if $f(1) = 0$, then for any positive integer n, we have $f(n) = f(1\cdot n) = f(1)\cdot f(n) = 0$. In other words, we are led to an arithmetic function that is identical to zero. Therefore, the restriction that $f(1) = 1$ is added to exclude the above less interesting function.

Analogously, we may replace the above multiplicative condition with an additive condition.

Definition 13.3 An arithmetic function f is

(i) *additive* if $f(mn) = f(m) + f(n)$ for all positive integers m and n with $(m,n) = 1$;

(ii) *completely additive* if $f(mn) = f(m) + f(n)$ for all positive integers m and n.

R For any additive function f, we always have $f(1)=0$.

We list here several simple but important arithmetic functions:

- the *constant function* $\mathbf{1}(n)$, defined by $\mathbf{1}(n)=1$ for all n — **completely multiplicative**;
- the *identity function* $\mathrm{id}(n)$, defined by $\mathrm{id}(n)=n$ for all n — **completely multiplicative**;
- the *unit function* $\varepsilon(n)$, defined by $\varepsilon(n)=1$ if $n=1$, and 0 otherwise — **completely multiplicative**;
- the function $\Omega(n)$, defined by the total number of prime factors of n (e.g. $\Omega(1)=0$, $\Omega(2)=1$, $\Omega(4)=2$, $\Omega(6)=2$, $\Omega(12)=3$, etc.) — **completely additive**;
- the function $\omega(n)$, defined by the number of distinct prime factors of n (e.g. $\omega(1)=0$, $\omega(2)=1$, $\omega(4)=1$, $\omega(6)=2$, $\omega(12)=2$, etc.) — **additive**.

13.2 Divisor functions

Definition 13.4 For s a given real or complex number, the *divisor function* $\sigma_s(n)$ is defined by

$$\sigma_s(n):=\sum_{d|n}d^s,$$

where the summation runs over all positive divisors of n. In particular, we define

$$d(n)=\sigma_0(n)=\sum_{d|n}1 \qquad \text{and} \qquad \sigma(n)=\sigma_1(n)=\sum_{d|n}d.$$

Theorem 13.1 Let $n=p_1^{\alpha_1}\cdots p_r^{\alpha_r}$ be in the canonical form. Then

$$d(n)=\prod_{k=1}^{r}(\alpha_k+1) \tag{13.1}$$

and for $s\neq 0$,

$$\sigma_s(n)=\prod_{k=1}^{r}\frac{p_k^{(\alpha_k+1)s}-1}{p_k^s-1}. \tag{13.2}$$

Proof. Noting that all divisors of n are of the form $p_1^{\beta_1}\cdots p_r^{\beta_r}$ with $0\le\beta_k\le\alpha_k$ for each k, we have

$$\sigma_s(n)=\sum_{d|n}d^s=\sum_{\beta_1=0}^{\alpha_1}\cdots\sum_{\beta_r=0}^{\alpha_r}\left(p_1^{\beta_1}\cdots p_r^{\beta_r}\right)^s=\prod_{k=1}^{r}\left(1+p_k^s+p_k^{2s}+\cdots+p_k^{\alpha_k s}\right).$$

We further get (13.1) and (13.2) by using the fact that $1+p^s+\cdots+p^{\alpha s}$ equals $\alpha+1$ if $s=0$, and $\frac{p^{(\alpha+1)s}-1}{p^s-1}$ if $s\neq 0$. ■

Corollary 13.2 For any s, the divisor function $\sigma_s(n)$ is multiplicative.

Proof. This is a direct implication of Theorem 13.1. ■

13.3 Möbius function

Recall that $\omega(n)$ counts the number of distinct prime factors of n.

Definition 13.5 An integer n is *squarefree* if no squares other than 1 divide n; otherwise, we say n is *squareful*.

▪ **Example 13.1** The first several positive squarefree integers are $1, 2, 3, 5, 6, 7, 10, 11, \dots$ and the first several positive squareful integers are $4, 8, 9, 12, 16, 18, 20, 24, \dots$ ▪

Definition 13.6 The *Möbius function* $\mu(n)$ is defined by

$$\mu(n) = \begin{cases} (-1)^{\omega(n)} & \text{if } n \text{ is squarefree,} \\ 0 & \text{otherwise.} \end{cases}$$

R The Möbius function was introduced by the German mathematician August Ferdinand Möbius (*J. Reine Angew. Math.* **9** (1832), 105–123).

▪ **Example 13.2** We have $\mu(1) = 1$, $\mu(2) = -1$, $\mu(3) = -1$, $\mu(4) = 0$, $\mu(5) = -1$, $\mu(6) = 1$, etc. ▪

Theorem 13.3 The Möbius function $\mu(n)$ is multiplicative.

Proof. First, we have $\mu(1) = 1$. Let us assume that m and n are such that $(m,n) = 1$. If one of m and n is squareful, then mn is also squareful, and hence $\mu(mn) = 0 = \mu(m)\mu(n)$. If both m and n are squarefree, so is mn as $(m,n) = 1$. Thus, $\mu(mn) = (-1)^{\omega(mn)} = (-1)^{\omega(m)+\omega(n)} = \mu(m)\mu(n)$ since $\omega(n)$ is additive. ■

Theorem 13.4 For $n \geq 1$,

$$\sum_{d|n} \mu(d) = \begin{cases} 1 & \text{if } n = 1, \\ 0 & \text{if } n > 1. \end{cases} \tag{13.3}$$

Proof. The formula is trivial when $n = 1$. For $n > 1$, we write n in the canonical form $n = p_1^{\alpha_1} \cdots p_r^{\alpha_r}$. Note that it suffices to consider squarefree divisors d of n in the sum $\sum_{d|n} \mu(d)$. We have

$$\begin{aligned} \sum_{d|n} \mu(d) &= \mu(1) + \mu(p_1) + \cdots + \mu(p_r) + \mu(p_1 p_2) + \cdots + \mu(p_{r-1} p_r) + \cdots + \mu(p_1 \cdots p_r) \\ &= \binom{r}{0} - \binom{r}{1} + \binom{r}{2} + \cdots + (-1)^r \binom{r}{r} \\ &= (1-1)^r = 0, \end{aligned}$$

as required. ■

R Recalling the definition of the unit function ε, we have

$$\varepsilon(n) = \sum_{d|n} \mu(d).$$

13.4 Euler's totient function revisited

Euler's totient function $\phi(n)$ was well studied in Sect. 4.2 and later lectures. In particular, we know that $\phi(n)$ is multiplicative. Also, we have shown in Theorem 4.5 that

$$\sum_{d|n}\phi(d)=n. \tag{13.4}$$

Now we establish a formula connecting Euler's totient function and the Möbius function.

Theorem 13.5 For $n\geq 1$,

$$\phi(n)=\sum_{d|n}\mu(d)\frac{n}{d}. \tag{13.5}$$

Proof. By the definition of $\phi(n)$, we have, with (13.3) applied, that

$$\phi(n)=\sum_{k=1}^{n}\varepsilon\big((k,n)\big)=\sum_{k=1}^{n}\sum_{d|(k,n)}\mu(d)=\sum_{k=1}^{n}\sum_{\substack{d|k\\ d|n}}\mu(d)=\sum_{d|n}\mu(d)\sum_{\substack{k=1\\ d|k}}^{n}1=\sum_{d|n}\mu(d)\frac{n}{d},$$

as required. ■

13.5 Mangoldt function

In this part, we introduce the *Mangoldt function* $\Lambda(n)$ which plays a crucial role in the study of the distribution of primes.

Definition 13.7 The *Mangoldt function* $\Lambda(n)$ is defined by

$$\Lambda(n)=\begin{cases}\log p & \text{if } n=p^{\alpha} \text{ with } p \text{ a prime and } \alpha \text{ a positive integer,}\\ 0 & \text{otherwise.}\end{cases}$$

R The Mangoldt function is named after the German mathematician Hans von Mangoldt.

Example 13.3 We have $\Lambda(1)=0$, $\Lambda(2)=\log 2$, $\Lambda(3)=\log 3$, $\Lambda(4)=\log 2$, $\Lambda(5)=\log 5$, $\Lambda(6)=0$, etc.

R The Mangoldt function $\Lambda(n)$ is neither multiplicative nor additive, for $\Lambda(6)\neq\Lambda(2)\Lambda(3)$ and $\Lambda(6)\neq\Lambda(2)+\Lambda(3)$.

Theorem 13.6 For $n\geq 1$,

$$\log n=\sum_{d|n}\Lambda(d). \tag{13.6}$$

Proof. This formula is trivial when $n=1$. For $n>1$, we write n in the canonical form $n=p_1^{\alpha_1}\cdots p_r^{\alpha_r}$. Then

$$\sum_{d|n}\Lambda(d)=\sum_{k=1}^{r}\big(\Lambda(p_k)+\Lambda(p_k^2)+\cdots+\Lambda(p_k^{\alpha_k})\big)=\sum_{k=1}^{r}\alpha_k\log p_k=\sum_{k=1}^{r}\log p_k^{\alpha_k}=\log n,$$

as desired. ■

Theorem 13.7 For $n \geq 1$,

$$\Lambda(n) = -\sum_{d|n} \mu(d) \log d. \tag{13.7}$$

Proof. The relation is trivial when $n = 1$. Also, if $n = p^\alpha$ with p a prime and α a positive integer, we have

$$-\sum_{d|p^\alpha} \mu(d) \log d = -\mu(1) \log 1 - \mu(p) \log p = \log p = \Lambda(p^\alpha).$$

Now we assume that n is written in the canonical form $n = p_1^{\alpha_1} \cdots p_r^{\alpha_r}$ with $r \geq 2$. Then

$$\begin{aligned}-\sum_{d|n} \mu(d) \log d = &\sum_{1 \leq i \leq r} \log p_i - \sum_{1 \leq i < j \leq r} \log p_i p_j \\ &+ \sum_{1 \leq i < j < k \leq r} \log p_i p_j p_k - \cdots + (-1)^{r-1} \log p_1 p_2 \cdots p_r.\end{aligned}$$

Note that $\log xy = \log x + \log y$. We then find that in the summation $\sum_{1 \leq i \leq r} \log p_i$, each $\log p_\ell$ appears $1 = \binom{r-1}{0}$ time; in the summation $\sum_{1 \leq i < j \leq r} \log p_i p_j$, each $\log p_\ell$ appears $r - 1 = \binom{r-1}{1}$ times; in the summation $\sum_{1 \leq i < j < k \leq r} \log p_i p_j p_k$, each $\log p_\ell$ appears $\binom{r-1}{2}$ times, etc. Hence,

$$\begin{aligned}-\sum_{d|n} \mu(d) \log d &= \sum_{\ell=1}^{r} \left(\binom{r-1}{0} - \binom{r-1}{1} + \binom{r-1}{2} - \cdots + (-1)^{r-1} \binom{r-1}{r-1} \right) \log p_\ell \\ &= \sum_{\ell=1}^{r} (1-1)^{r-1} \log p_\ell = 0.\end{aligned}$$

However, for $n = p_1^{\alpha_1} \cdots p_r^{\alpha_r}$ with $r \geq 2$, we also have $\Lambda(n) = 0$ by definition. The desired identity holds. ■

Corollary 13.8 For $n \geq 1$,

$$\Lambda(n) = \sum_{d|n} \mu(d) \log \frac{n}{d}. \tag{13.8}$$

Proof. Note that

$$\sum_{d|n} \mu(d) \log \frac{n}{d} = \sum_{d|n} \mu(d) \big(\log n - \log d \big) = (\log n) \sum_{d|n} \mu(d) - \sum_{d|n} \mu(d) \log d.$$

Since $(\log n) \sum_{d|n} \mu(d) = (\log n) \cdot \varepsilon(n) = 0$ for every $n \geq 1$, we arrive at the required result by recalling (13.7). ■

14. Möbius inversion formula

14.1 Möbius inversion formula

The pair of relations (13.4) and (13.5), and the pair of relations (13.6) and (13.8) are indeed special cases of a general phenomenon, known as the *Möbius inversion*.

Theorem 14.1 (Möbius Inversion Formula). Let $f(n)$ and $g(n)$ be arithmetic functions. If

$$g(n) = \sum_{d|n} f(d) \tag{14.1}$$

then

$$f(n) = \sum_{d|n} \mu(d) g\left(\frac{n}{d}\right), \tag{14.2}$$

and vice versa.

R In (13.4) and (13.5), we have $f = \phi$ and $g = \mathrm{id}$; in (13.6) and (13.8), we have $f = \Lambda$ and $g = \log$.

Proof. We first prove (14.2) by (14.1). Note that

$$\begin{aligned}\sum_{d|n} \mu(d) g\left(\frac{n}{d}\right) &= \sum_{d|n} \mu(d) \sum_{d'|\frac{n}{d}} f(d') = \sum_{\substack{d,d'\\ dd'|n}} \mu(d) f(d') \\ &= \sum_{d'|n} f(d') \sum_{d|\frac{n}{d'}} \mu(d) = \sum_{d'|n} f(d') \varepsilon\left(\frac{n}{d'}\right) = f(n),\end{aligned}$$

where we make use of (13.3). Conversely, to show (14.1) from (14.2), we first require the trivial fact that for any arithmetic function $a(n)$,

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

Rewriting (14.2) as

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

it follows that

$$\begin{aligned}\sum_{d|n} f(d) &= \sum_{d|n} f\left(\frac{n}{d}\right) = \sum_{d|n}\sum_{d'|\frac{n}{d}} \mu\left(\frac{n/d}{d'}\right) g(d') = \sum_{\substack{d,d'\\ dd'|n}} \mu\left(\frac{n}{dd'}\right) g(d')\\ &= \sum_{d'|n} g(d') \sum_{d|\frac{n}{d'}} \mu\left(\frac{n/d'}{d}\right) = \sum_{d'|n} g(d') \sum_{d|\frac{n}{d'}} \mu(d) = \sum_{d'|n} g(d')\varepsilon\left(\frac{n}{d'}\right) = g(n),\end{aligned}$$

where (13.3) is also applied. ■

There is a slightly different type of Möbius inversion formula working for functions defined on real $x > 0$. Below, in the summation $\sum_{n\le x}$, the index n runs over all positive integers no larger than x.

Theorem 14.2 Let $F(x)$ and $G(x)$ be functions defined on real $x > 0$. If

$$G(x) = \sum_{n\le x} F\left(\frac{x}{n}\right) \tag{14.3}$$

then

$$F(x) = \sum_{n\le x} \mu(n) G\left(\frac{x}{n}\right), \tag{14.4}$$

and vice versa.

Proof. We first prove (14.4) by (14.3). Note that

$$\begin{aligned}\sum_{n\le x} \mu(n) G\left(\frac{x}{n}\right) &= \sum_{n\le x} \mu(n) \sum_{m\le \frac{x}{n}} F\left(\frac{x/n}{m}\right) = \sum_{\substack{m,n\\ mn\le x}} \mu(n) F\left(\frac{x}{mn}\right)\\ (\text{with } N=mn) &= \sum_{N\le x} F\left(\frac{x}{N}\right) \sum_{n|N} \mu(n) = \sum_{N\le x} F\left(\frac{x}{N}\right)\varepsilon(N) = F(x).\end{aligned}$$

Conversely, to show (14.3) from (14.4), we have

$$\begin{aligned}\sum_{n\le x} F\left(\frac{x}{n}\right) &= \sum_{n\le x}\sum_{m\le \frac{x}{n}} \mu(m) G\left(\frac{x/n}{m}\right) = \sum_{\substack{m,n\\ mn\le x}} \mu(m) G\left(\frac{x}{mn}\right)\\ (\text{with } N=mn) &= \sum_{N\le x} G\left(\frac{x}{N}\right) \sum_{m|N} \mu(m) = \sum_{N\le x} G\left(\frac{x}{N}\right)\varepsilon(N) = G(x),\end{aligned}$$

as required. ■

14.2 Multiplicative Möbius inversion formula

Another important variant of Möbius inversion formula is in the multiplicative notation.

Theorem 14.3 Let $f(n)$ and $g(n)$ be arithmetic functions such that $f(n) \ne 0$ and $g(n) \ne 0$ for all n. If

$$g(n) = \prod_{d|n} f(d) \tag{14.5}$$

then

$$f(n) = \prod_{d|n} g\left(\frac{n}{d}\right)^{\mu(d)}, \tag{14.6}$$

and vice versa.

Proof. We first prove (14.6) by (14.5). Note that

$$\begin{aligned}\prod_{d|n} g\left(\frac{n}{d}\right)^{\mu(d)} &= \prod_{d|n}\left(\prod_{d'|\frac{n}{d}} f(d')\right)^{\mu(d)} = \prod_{d|n}\prod_{d'|\frac{n}{d}} f(d')^{\mu(d)} = \prod_{d'|n}\prod_{d|\frac{n}{d'}} f(d')^{\mu(d)}\\ &= \prod_{d'|n} f(d')^{\sum_{d|\frac{n}{d'}}\mu(d)} = \prod_{d'|n} f(d')^{\varepsilon(n/d')} = f(n).\end{aligned}$$

Conversely, to show (14.5) from (14.6), we have

$$\begin{aligned}\prod_{d|n} f(d) = \prod_{d|n} f\left(\frac{n}{d}\right) &= \prod_{d|n}\prod_{d'|\frac{n}{d}} g(d')^{\mu\left(\frac{n/d}{d'}\right)} = \prod_{d'|n}\prod_{d|\frac{n}{d'}} g(d')^{\mu\left(\frac{n}{dd'}\right)}\\ &= \prod_{d'|n} g(d')^{\sum_{d|\frac{n}{d'}}\mu\left(\frac{n/d'}{d}\right)} = \prod_{d'|n} g(d')^{\sum_{d|\frac{n}{d'}}\mu(d)} = \prod_{d'|n} g(d')^{\varepsilon(n/d')} = g(n),\end{aligned}$$

as required. ■

R Intuitively, for positive-valued f and g, we may define $\tilde{f}(n) = \log f(n)$ and $\tilde{g}(n) = \log g(n)$. By taking logarithm in (14.5) and (14.6), their equivalence becomes

$$\tilde{g}(n) = \sum_{d|n} \tilde{f}(d) \qquad \Longleftrightarrow \qquad \tilde{f}(n) = \sum_{d|n} \mu(d)\tilde{g}\left(\frac{n}{d}\right),$$

which is exactly the usual Möbius inversion formula.

14.3 Dirichlet convolutions

The Möbius inversion formula can be further understood in a more abstract way, through *Dirichlet convolutions*, named after the German mathematician Peter Gustav Lejeune Dirichlet.

Definition 14.1 For arithmetic functions f and g, their *Dirichlet convolution* is defined to be an arithmetic function h with

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

where the summation runs over all positive divisors of n. We write

$$h = f * g.$$

The Dirichlet convolution satisfies the following algebraic properties.

Theorem 14.4 For any arithmetic functions u, v and w, we have

(i) $u * v = v * u$ (commutative law);

(ii) $(u * v) * w = u * (v * w)$ (associative law).

Proof. It is straightforward to verify that

$$(u*v)(n) = (v*u)(n) = \sum_{\substack{a,b \\ ab=n}} u(a)v(b)$$

and

$$\big((u*v)*w\big)(n) = \big(u*(v*w)\big)(n) = \sum_{\substack{a,b,c \\ abc=n}} u(a)v(b)w(c),$$

where a, b and c run over positive integers. ■

Theorem 14.5 Let ε be the unit function. For any arithmetic function f, we have $f*\varepsilon = \varepsilon * f = f$.

Proof. We have

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

as required. ■

Theorem 14.6 Let f be an arithmetic function with $f(1) \neq 0$. Then there exists a unique arithmetic function g such that $f*g = g*f = \varepsilon$. Moreover, g is given by

$$g(1) = \frac{1}{f(1)} \tag{14.7}$$

and for $n \geq 2$,

$$g(n) = -\frac{1}{f(1)} \sum_{\substack{d|n \\ d<n}} f\left(\frac{n}{d}\right) g(d). \tag{14.8}$$

Proof. First, we note that $(f*g)(1) = f(1)g(1) = \varepsilon(1) = 1$ gives $g(1) = 1/f(1)$. For $n \geq 2$, we have $\varepsilon(n) = 0$, and hence,

$$0 = (f*g)(n) = (g*f)(n) = \sum_{d|n} f\left(\frac{n}{d}\right) g(d) = f(1)g(n) + \sum_{\substack{d|n \\ d<n}} f\left(\frac{n}{d}\right) g(d).$$

Hence, we may iteratively determine the unique $g(n)$ by (14.8). ■

Definition 14.2 Given an arithmetic function f with $f(1) \neq 0$, we call the unique arithmetic function g such that $f*g = g*f = \varepsilon$ the *Dirichlet inverse* of f, denoted by $g = f^{-1}$.

Theorem 14.7 For any arithmetic functions with $f(1) \neq 0$ and $g(1) \neq 0$, we have $(f*g)^{-1} = f^{-1} * g^{-1}$.

Proof. We have $(f*g)*(f^{-1}*g^{-1}) = (f*f^{-1})*(g*g^{-1}) = \varepsilon * \varepsilon = \varepsilon$, as required. ■

R In the language of group theory, the set of arithmetic functions f with $f(1) \neq 0$ forms an abelian group under the operation “$*$” (Dirichlet convolution), and the identity element of this group is the unit function ε.

Corollary 14.8 The Möbius function μ and the constant function $\mathbf{1}$ are Dirichlet inverses of one another.

Proof. We simply rewrite the relation (13.3), $\sum_{d\mid n}\mu(d)=\varepsilon(n)$, in terms of the Dirichlet convolution, and find that $\mu*\mathbf{1}=\varepsilon$, thereby yielding the desired result. ■

R We may also interpret the Möbius inversion formula in this setting by noting that the Möbius inversion is exactly the equivalence

$$g=f*\mathbf{1}\qquad\Longleftrightarrow\qquad f=g*\mu.$$

This is trivial since if $g=f*\mathbf{1}$, then $g*\mu=(f*\mathbf{1})*\mu=f*(\mu*\mathbf{1})=f*\varepsilon=f$; and if $f=g*\mu$, then $f*\mathbf{1}=(g*\mu)*\mathbf{1}=g*(\mu*\mathbf{1})=g*\varepsilon=g$.

Now we consider Dirichlet convolutions on multiplicative functions.

Theorem 14.9 If f and g are multiplicative functions, so is their Dirichlet convolution $f*g$.

Proof. We write $h=f*g$. Let m and n be positive integers with $(m,n)=1$. We use the fact that if $d\mid mn$, then we may uniquely write $d=ab$ with $a\mid m$ and $b\mid n$. In particular, $(a,b)=1$ and $(\frac{m}{a},\frac{n}{b})=1$. Now,

$$\begin{aligned}h(mn)&=\sum_{d\mid mn}f(d)g\left(\frac{mn}{d}\right)=\sum_{a\mid m,b\mid n}f(ab)g\left(\frac{mn}{ab}\right)=\sum_{a\mid m,b\mid n}f(a)f(b)g\left(\frac{m}{a}\right)g\left(\frac{n}{b}\right)\\&=\sum_{a\mid m}f(a)g\left(\frac{m}{a}\right)\sum_{b\mid n}f(b)g\left(\frac{n}{b}\right)=h(m)h(n).\end{aligned}$$

Hence, $h=f*g$ is multiplicative. ■

Theorem 14.10 If f is a multiplicative function, so is its Dirichlet inverse f^{-1}.

Proof. Noting that f is multiplicative, we have $f(1)=1$, and hence $f^{-1}(1)=\frac{1}{f(1)}=1$. Now we shall show that for every positive integer N, $f^{-1}(N)=f^{-1}(m)f^{-1}(n)$ holds for any positive integers m and n with $(m,n)=1$ and $mn=N$. We argue by induction on N. The base case $N=1$ is confirmed by the fact that $f^{-1}(1)=1$. Assume that the claim is true for $1,\ldots,N-1$ with $N\geq 2$, and we shall prove the case of N. Note that

$$\begin{aligned}\varepsilon(N)=(f^{-1}*f)(mn)&=\sum_{a\mid m,b\mid n}f^{-1}(ab)f\left(\frac{mn}{ab}\right)\\&=f^{-1}(mn)f(1)+\sum_{\substack{a\mid m,b\mid n\\ab<N}}f^{-1}(ab)f\left(\frac{mn}{ab}\right)\\\text{(induc. assump.)}&=f^{-1}(mn)f(1)+\sum_{\substack{a\mid m,b\mid n\\ab<N}}f^{-1}(a)f^{-1}(b)f\left(\frac{m}{a}\right)f\left(\frac{n}{b}\right)\\&=f^{-1}(mn)f(1)-f^{-1}(m)f^{-1}(n)f(1)f(1)+\sum_{a\mid m,b\mid n}f^{-1}(a)f^{-1}(b)f\left(\frac{m}{a}\right)f\left(\frac{n}{b}\right)\\&=f^{-1}(N)-f^{-1}(m)f^{-1}(n)+(f^{-1}*f)(m)(f^{-1}*f)(n)\\&=f^{-1}(N)-f^{-1}(m)f^{-1}(n)+\varepsilon(N),\end{aligned}$$

thereby implying that $f^{-1}(N)=f^{-1}(m)f^{-1}(n)$, as required. ■

R The set of multiplicative functions forms a subgroup of the group of all arithmetic functions f with $f(1) \neq 0$ under the Dirichlet convolution.

14.4 Ramanujan's sums

We first adopt a conventional notation in analytic number theory.

Definition 14.3 For any complex number τ, we define

$$e(\tau) := e^{2\pi i\tau}.$$

A trivial fact about this function is that for any integer k,

$$e(\tau + k) = e(\tau), \tag{14.9}$$

since $e^{2\pi i(\tau+k)} = e^{2\pi i\tau} \cdot e^{2k\pi i} = e^{2\pi i\tau}$.

Now we introduce *Ramanujan's sums*, which are crucial in, for instance, the proof of I. M. Vinogradov's theorem (*Recueil Math.* **2** (1937), 179–195) that *every sufficiently large odd number is the sum of three primes.*

Definition 14.4 For q and n positive integers, *Ramanujan's sums* are defined by

$$c_q(n) := \sum_{\substack{1 \le a \le q \\ (a,q)=1}} e\left(\frac{an}{q}\right).$$

R Ramanujan's sums were introduced by Ramanujan (*Trans. Cambridge Philos. Soc.* **22** (1918), no. 13, 259–276).

We introduce another sum for q and n positive integers:

$$\eta_q(n) := \sum_{1 \le a \le q} e\left(\frac{an}{q}\right).$$

Lemma 14.11 For positive integers q and n,

$$\eta_q(n) = \begin{cases} q & \text{if } q \mid n, \\ 0 & \text{if } q \nmid n. \end{cases} \tag{14.10}$$

In particular, for positive integers s and t with $(s,t) = 1$, we have $\eta_s(n)\eta_t(n) = \eta_{st}(n)$.

Proof. Let $d = (q,n)$, and write $q = q'd$ and $n = n'd$. Noting that $(q',n') = 1$, we have $\{an' : 1 \le a \le q'\}$ forms a complete system modulo q'. Now,

$$\eta_q(n) = \sum_{1 \le a \le q} e\left(\frac{an}{q}\right) = \sum_{1 \le a \le q'd} e\left(\frac{an'}{q'}\right) = d \sum_{1 \le a \le q'} e\left(\frac{an'}{q'}\right) = d \sum_{1 \le a \le q'} e\left(\frac{a}{q'}\right),$$

where we use (14.9) in the second last equality. Note that

$$\sum_{1 \le a \le q'} e\left(\frac{a}{q'}\right) = \begin{cases} 1 & \text{if } q' = 1, \\ 0 & \text{if } q' > 1. \end{cases}$$

Finally, we use the fact that $q' = 1$ if and only if $q = d = (q,n)$, namely, $q \mid n$, as desired. The second part is a direct consequence of (14.10). ■

Now we establish a relation between $c_q(n)$ and $\eta_q(n)$.

Theorem 14.12 For positive integers q and n,

$$\eta_q(n) = \sum_{d|q} c_d(n). \tag{14.11}$$

Proof. We use the fact that $\{\frac{a}{q} : 1 \le a \le q\} = \cup_{d|q}\{\frac{b}{d} : 1 \le b \le d \text{ and } (b,d) = 1\}$, by simplifying each $\frac{a}{q}$ to its irreducible expression. Hence,

$$\sum_{1\le a\le q} e\left(\frac{an}{q}\right) = \sum_{d|q} \sum_{\substack{1\le b\le d\\(b,d)=1}} e\left(\frac{bn}{d}\right),$$

as required. ■

Let us treat $\eta_q(n)$ and $c_q(n)$ as functions in q with n fixed, and define $H(q) := \eta_q(n)$ and $C(q) := c_q(n)$ for clarity. Then we may paraphrase (14.11) as

$$H = C * \mathbf{1}, \tag{14.12}$$

and equivalently,

$$C = H * \mu. \tag{14.13}$$

Corollary 14.13 Let n be a positive integer. For positive integers s and t with $(s,t) = 1$,

$$c_s(n)c_t(n) = c_{st}(n). \tag{14.14}$$

Proof. We use Theorem 14.9 by noting that both H and μ are multiplicative. ■

Corollary 14.14 For positive integers q and n,

$$c_q(n) = \sum_{d|q,d|n} \mu\left(\frac{q}{d}\right) d. \tag{14.15}$$

Proof. Note that (14.13) can be explicitly written as

$$c_q(n) = \sum_{d|q} \mu\left(\frac{q}{d}\right) \eta_d(n).$$

The desired relation follows with recourse to (14.10). ■

Theorem 14.15 For positive integers q and n,

$$c_q(n) = \mu\left(\frac{q}{(q,n)}\right) \frac{\phi(q)}{\phi\left(\frac{q}{(q,n)}\right)}. \tag{14.16}$$

Proof. For convenience, we write

$$R_q(n) := \mu\left(\frac{q}{(q,n)}\right) \frac{\phi(q)}{\phi\left(\frac{q}{(q,n)}\right)}. \tag{14.17}$$

Let n be an arbitrary positive integer. Note that $c_1(n) = R_1(n)$. Also, let s and t be such that $(s,t) = 1$. Then $(st,n) = (s,n) \cdot (t,n)$ and $\left(\frac{s}{(s,n)}, \frac{t}{(t,n)}\right) = 1$. Thus,

$$R_{st}(n) = \mu\left(\frac{st}{(st,n)}\right) \frac{\phi(st)}{\phi\left(\frac{st}{(st,n)}\right)} = \mu\left(\frac{s}{(s,n)}\right) \mu\left(\frac{t}{(t,n)}\right) \frac{\phi(s)\phi(t)}{\phi\left(\frac{s}{(s,n)}\right)\phi\left(\frac{t}{(t,n)}\right)} = R_s(n)R_t(n).$$

Recalling (14.14), it suffices to prove for prime powers p^α that $c_{p^\alpha}(n) = R_{p^\alpha}(n)$. Finally, it is straightforward to compute from (14.15) and (14.17) that

$$c_{p^\alpha}(n) = R_{p^\alpha}(n) = \begin{cases} p^\alpha - p^{\alpha-1} & \text{if } (p^\alpha, n) = p^\alpha, \\ -p^{\alpha-1} & \text{if } (p^\alpha, n) = p^{\alpha-1}, \\ 0 & \text{otherwise.} \end{cases}$$

The desired relation holds. ■

15. Average of arithmetic functions

15.1 Asymptotic relations

Given an arithmetic function f, one of the basic problems in analytic number theory concerns the asymptotic analysis of the partial sum

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

where the summation runs over all **positive integers** no larger than x. Meanwhile, we are also often interested in the behavior of

$$\sum_{p\le x} f(p)$$

in which the index p means that we are summing over **primes** no larger than x.

To begin with, we introduce some useful notations for asymptotic analysis.

Definition 15.1 (Bachmann–Landau Notations).

▷ The *big O notation* $f(x)=O(g(x))$ means that there exists a constant C such that $|f(x)|\le C|g(x)|$;

▷ The *small o notation* $f(x)=o(g(x))$ means that $\lim f(x)/g(x)=0$.

R Big O and small o belong to a family of notations invented by the German mathematicians Paul Bachmann and Edmund Landau.

Definition 15.2 (Vinogradov Notations).

▷ The notation $f(x)\ll g(x)$ means that $f(x)=O(g(x))$;

▷ The notation $f(x)\gg g(x)$ means that $g(x)\ll f(x)$.

R The two notations were introduced by the Russian mathematician Ivan Matveevich Vinogradov.

Definition 15.3

▷ The *asymptotic equivalence* symbol $f(x)\sim g(x)$ means that $\lim f(x)/g(x)=1$;

▷ The *order of magnitude estimate* symbol $f(x)\asymp g(x)$ means that both $f(x)\ll g(x)$ and $g(x)\ll f(x)$ hold. Equivalently, there exist constants C_1 and C_2 such that

$$C_1|g(x)| \leq |f(x)| \leq C_2|g(x)|.$$

15.2 Abel's summation formula

On many occasions, a partial sum can be nicely estimated by comparing it with an integral. To do so, a summation formula due to the Norwegian mathematician Niels Henrik Abel, and especially its special case that was obtained earlier by Euler, play a crucial role.

Definition 15.4 We denote by $\lfloor x \rfloor$ the largest integer not exceeding x, and by $\{x\} := x - \lfloor x \rfloor$.

Theorem 15.1 (Abel's Summation Formula). Let $a : \mathbb{Z}_{>0} \to \mathbb{C}$ be an arithmetic function, let $0 < y < x$ be real numbers, and let $f : [y,x] \to \mathbb{C}$ be a function with continuous derivative f' on the interval $[y,x]$. Then

$$\sum_{y<n\leq x} a(n)f(n) = A(x)f(x) - A(y)f(y) - \int_y^x A(t)f'(t)dt, \tag{15.1}$$

where $A(t) = \sum_{n\leq t} a(n)$.

Proof. We start by observing that $A(t) = A(\lfloor t \rfloor)$ and $A(t+1) - A(t) = a(\lfloor t \rfloor + 1)$. It is also straightforward to see that if there is no integer in the interval $(y,x]$, both sides of (15.1) are zero. Now we assume that there is at least one integer in $(y,x]$, and evaluate the integral on the right-hand side of (15.1):

$$\begin{aligned}
\int_y^x A(t)f'(t)dt &= \left(\int_y^{\lfloor y \rfloor+1} + \int_{\lfloor y \rfloor+1}^{\lfloor y \rfloor+2} + \cdots + \int_{\lfloor x \rfloor-1}^{\lfloor x \rfloor} + \int_{\lfloor x \rfloor}^{x}\right) A(t)f'(t)dt \\
&= A(y)\big(f(\lfloor y \rfloor+1) - f(y)\big) + A(y+1)\big(f(\lfloor y \rfloor+2) - f(\lfloor y \rfloor+1)\big) + \cdots \\
&\quad + A(x-1)\big(f(\lfloor x \rfloor) - f(\lfloor x \rfloor - 1)\big) + A(x)\big(f(x) - f(\lfloor x \rfloor)\big) \\
&= A(x)f(x) - A(y)f(y) - \big(A(y+1) - A(y)\big)f(\lfloor y \rfloor+1) - \cdots \\
&\quad - \big(A(x) - A(x-1)\big)f(\lfloor x \rfloor) \\
&= A(x)f(x) - A(y)f(y) - a(\lfloor y \rfloor+1)f(\lfloor y \rfloor+1) - \cdots - a(\lfloor x \rfloor)f(\lfloor x \rfloor) \\
&= A(x)f(x) - A(y)f(y) - \sum_{y<n\leq x} a(n)f(n),
\end{aligned}$$

as required. ■

R A more advanced way to think of Abel's summation formula is by means of the Riemann–Stieltjes integral:

$$\begin{aligned}
\sum_{y<n\leq x} a(n)f(n) &= \int_y^x f(t)dA(t) \\
&= f(x)A(x) - f(y)A(y) - \int_y^x A(t)df(t),
\end{aligned}$$

where we use integration by parts for the second equality.

It is particularly useful to choose $a(n) = 1$ for all n in Abel's summation formula, and then observe that

$$A(t) = \sum_{n\leq t} 1 = \lfloor t \rfloor.$$

We may recover a summation formula due to Euler.

Corollary 15.2 (Euler's Summation Formula). Let $0 < y < x$ be real numbers, and let $f : [y,x] \to \mathbb{C}$ be a function with continuous derivative f' on the interval $[y,x]$. Then

$$\sum_{y<n\leq x} f(n) = \int_y^x f(t)dt + \int_y^x \{t\}f'(t)dt + \{y\}f(y) - \{x\}f(x). \tag{15.2}$$

Proof. By choosing $a(n) = 1$ for all n in (15.1), we have

$$\sum_{y<n\leq x} f(n) = \lfloor x \rfloor f(x) - \lfloor y \rfloor f(y) - \int_y^x \lfloor t \rfloor f'(t)dt.$$

Also, it follows from integration by parts that

$$\int_y^x f(t)dt = f(x) - f(y) - \int_y^x tf'(t)dt.$$

Combining the above two relations gives (15.2) by recalling that $\{x\} = x - \lfloor x \rfloor$. ■

In the sequel, we present some applications of Euler's summation formula. Here,

$$\gamma := \lim_{x\to\infty}\left(\sum_{n\leq x}\frac{1}{n} - \log x\right) = 1 - \int_1^\infty \frac{\{t\}}{t^2}dt = 0.577215\cdots$$

is the *Euler–Mascheroni constant*, named after Euler and the Italian mathematician Lorenzo Mascheroni;

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

with s a complex number such that $\Re(s) > 1$ is the *Riemann zeta function* which is absolutely convergent in this half-plane.

Theorem 15.3 As $x \to \infty$,

(i) $\displaystyle\sum_{n\leq x}\frac{1}{n} = \log x + \gamma + O(x^{-1})$;

(ii) $\displaystyle\sum_{n\leq x}\frac{1}{n^s} = \zeta(s) + O(x^{1-s})$ if $\Re(s) > 1$;

(iii) $\displaystyle\sum_{n\leq x}\frac{1}{n^s} = \frac{x^{1-s}}{1-s} + O(1)$ if $0 < \Re(s) \leq 1$ and $s \neq 1$;

(iv) $\displaystyle\sum_{n\leq x} n^\alpha = \frac{x^{\alpha+1}}{\alpha+1} + O(x^\alpha)$ if $\Re(\alpha) \geq 0$.

R Parts (ii) and (iii) can be improved uniformly. It is known that the Riemann zeta function has an analytic continuation to $\mathbb{C}\backslash\{1\}$. In particular, we will show in Theorem 18.2 that, for $s \neq 1$ with $0 < \Re(s) \leq 1$, $\zeta(s)$ is continued analytically as

$$\zeta(s) = \frac{s}{s-1} - s\int_1^\infty \frac{\{t\}}{t^{s+1}}dt.$$

Mimicking the proof of Part (iii) with Euler's summation formula applied with $f(t) = t^{-s}$ for all complex $s \neq 1$ with $\Re(s) > 0$, we have

$$\sum_{n\leq x}\frac{1}{n^s} = \frac{x^{1-s}}{1-s} + \zeta(s) + O(x^{-s}).$$

This is left as an exercise.

Proof. **(i).** We take $f(t)=t^{-1}$ in Euler's summation formula and find that

$$\sum_{n\le x}\frac{1}{n}=\int_{1^-}^{x}\frac{dt}{t}-\int_{1^-}^{x}\frac{\{t\}}{t^2}dt+1-\frac{\{x\}}{x}=\log x-\int_{1}^{x}\frac{\{t\}}{t^2}dt+1+O(x^{-1})$$
$$=\log x+1-\int_{1}^{\infty}\frac{\{t\}}{t^2}dt+\int_{x}^{\infty}\frac{\{t\}}{t^2}dt+O(x^{-1})=\log x+\gamma+O(x^{-1}),$$

since $\int_x^\infty\frac{\{t\}}{t^2}dt\ll\int_x^\infty\frac{1}{t^2}dt=x^{-1}$.

(ii). We directly note that, for $\Re(s)>1$,

$$\sum_{n\le x}\frac{1}{n^s}=\zeta(s)-\sum_{n>x}\frac{1}{n^s}=\zeta(s)+O\left(\int_x^\infty\frac{dt}{t^s}\right)=\zeta(s)+O(x^{1-s}).$$

(iii). With $f(t)=t^{-s}$ where $0<\Re(s)\le 1$ and $s\ne 1$, we know from Euler's summation formula that

$$\sum_{n\le x}\frac{1}{n^s}=\int_{1^-}^{x}\frac{dt}{t^s}-s\int_{1^-}^{x}\frac{\{t\}}{t^{s+1}}dt+1-\frac{\{x\}}{x^s}=\int_{1}^{x}\frac{dt}{t^s}+O(1)=\frac{x^{1-s}}{1-s}+O(1).$$

(iv). With $f(t)=t^{\alpha}$ where $\Re(\alpha)\ge 0$, Euler's summation formula gives us that

$$\sum_{n\le x}n^{\alpha}=\int_{1^-}^{x}t^{\alpha}dt+\alpha\int_{1^-}^{x}\{t\}t^{\alpha-1}dt+1-\{x\}x^{\alpha}=\int_{1}^{x}t^{\alpha}dt+O(x^{\alpha})=\frac{x^{\alpha+1}}{\alpha+1}+O(x^{\alpha}),$$

as required. ■

15.3 Average order of $\sigma(n)$

Theorem 15.4 As $x\to\infty$,

$$\sum_{n\le x}\sigma(n)=\frac{\zeta(2)}{2}x^2+O(x\log x). \tag{15.3}$$

Proof. We have

$$\sum_{n\le x}\sigma(n)=\sum_{n\le x}\sum_{m|n}m=\sum_{\substack{m,d\\ md\le x}}m=\sum_{d\le x}\sum_{m\le\frac{x}{d}}m=\sum_{d\le x}\frac{1}{2}\left\lfloor\frac{x}{d}\right\rfloor\left(\left\lfloor\frac{x}{d}\right\rfloor+1\right)$$
$$=\frac{1}{2}\sum_{d\le x}\left(\frac{x}{d}\right)^2+O\left(\sum_{d\le x}\frac{x}{d}\right)=\frac{\zeta(2)}{2}x^2+O(x\log x),$$

where we make use of Theorem 15.3, Parts (i) and (ii). ■

Theorem 15.5 Let $\alpha\ne 1$ be a complex number with $\Re(\alpha)>0$. As $x\to\infty$,

$$\sum_{n\le x}\sigma_{\alpha}(n)=\frac{\zeta(\alpha+1)}{\alpha+1}x^{\alpha+1}+O(x^{\max\{1,\Re(\alpha)\}}). \tag{15.4}$$

Proof. We have

$$\sum_{n\le x}\sigma_{\alpha}(n)=\sum_{n\le x}\sum_{m|n}m^{\alpha}=\sum_{\substack{m,d\\ md\le x}}m^{\alpha}=\sum_{d\le x}\sum_{m\le\frac{x}{d}}m^{\alpha}=\sum_{d\le x}\left(\frac{(\frac{x}{d})^{\alpha+1}}{\alpha+1}+O\left(\left(\frac{x}{d}\right)^{\alpha}\right)\right)$$

$$= \frac{x^{\alpha+1}}{\alpha+1}\sum_{d\le x}\frac{1}{d^{\alpha+1}} + O\left(\sum_{d\le x}\frac{x^{\alpha}}{d^{\alpha}}\right) = \left(\frac{\zeta(\alpha+1)}{\alpha+1}x^{\alpha+1}+O(x)\right)+O(x^{\max\{1,\Re(\alpha)\}}),$$

where we make use of Theorem 15.3, Parts (ii), (iii) and (iv). ■

15.4 Average order of $\phi(n)$

Theorem 15.6 As $x\to\infty$,

$$\sum_{n\le x}\phi(n) = \frac{1}{2\zeta(2)}x^2 + O(x\log x). \tag{15.5}$$

Proof. We recall (13.5) and obtain

$$\sum_{n\le x}\phi(n) = \sum_{n\le x}\sum_{d\mid n}\mu(d)\frac{n}{d} = \sum_{\substack{m,d\\ md\le x}}\mu(d)m = \sum_{d\le x}\mu(d)\sum_{m\le \frac{x}{d}} m = \sum_{d\le x}\frac{\mu(d)}{2}\left(\left(\frac{x}{d}\right)^2 + O\left(\frac{x}{d}\right)\right)$$

$$= \frac{x^2}{2}\sum_{d\le x}\frac{\mu(d)}{d^2} + O\left(\sum_{d\le x}\frac{x}{d}\right) = \frac{x^2}{2}\sum_{d\le x}\frac{\mu(d)}{d^2} + O(x\log x).$$

Finally, we will show later in Example 16.4 that

$$\sum_{d\ge 1}\frac{\mu(d)}{d^2} = \frac{1}{\zeta(2)}.$$

Hence,

$$\frac{x^2}{2}\sum_{d\le x}\frac{\mu(d)}{d^2} = \frac{x^2}{2}\left(\sum_{d\ge 1}\frac{\mu(d)}{d^2} - \sum_{d>x}\frac{\mu(d)}{d^2}\right) = \frac{x^2}{2}\sum_{d\ge 1}\frac{\mu(d)}{d^2} + O\left(x^2\int_x^\infty\frac{dt}{t^2}\right) = \frac{x^2}{2\zeta(2)} + O(x),$$

thereby confirming the desired relation. ■

15.5 Dirichlet hyperbola method

For the purpose of getting a better estimate of the partial sum of the Dirichlet convolution of certain arithmetic functions, we sometimes require a trick due to Dirichlet, known as the *Dirichlet hyperbola method.*

Theorem 15.7 (Dirichlet Hyperbola Method). Let f and g be arithmetic functions and define

$$F(x) = \sum_{n\le x} f(n) \qquad \text{and} \qquad G(x) = \sum_{n\le x} g(n).$$

Then for any $1\le M\le x$,

$$\sum_{n\le x}(f*g)(n) = \sum_{u\le M} f(u)G\left(\frac{x}{u}\right) + \sum_{v\le x/M} g(v)F\left(\frac{x}{v}\right) - F(M)G\left(\frac{x}{M}\right). \tag{15.6}$$

Proof. We have

$$\sum_{n\le x}(f*g)(n) = \sum_{\substack{u,v\\ uv\le x}} f(u)g(v).$$

Now we consider the set of lattices $S = \{(u,v) \in \mathbb{Z}^2_{>0} : uv \leq x\}$. By the inclusion-exclusion principle, we may rewrite S as $S = S_{\mathsf{L}} \cup S_{\mathsf{B}} \backslash S_{\mathsf{O}}$, where

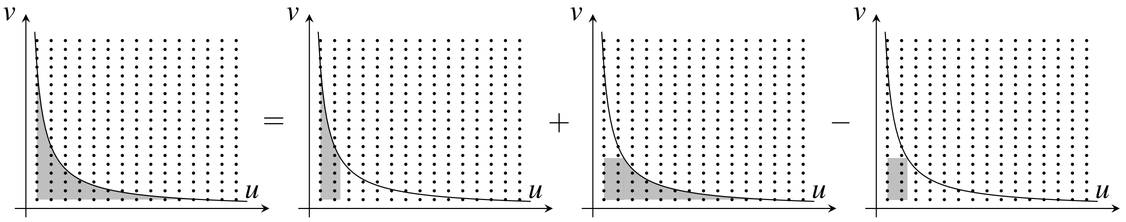
$$\begin{aligned} S_{\mathsf{Left}} &:= \{(u,v) \in \mathbb{Z}^2_{>0} : uv \leq x \text{ and } u \leq M\}, \\ S_{\mathsf{Below}} &:= \{(u,v) \in \mathbb{Z}^2_{>0} : uv \leq x \text{ and } v \leq x/M\}, \\ S_{\mathsf{Overlapping}} &:= \{(u,v) \in \mathbb{Z}^2_{>0} : u \leq M \text{ and } v \leq x/M\}. \end{aligned}$$

Hence,

$$\sum_{n \leq x} (f * g)(n) = \sum_{\substack{u \leq M \\ uv \leq x}} f(u)g(v) + \sum_{\substack{v \leq x/M \\ uv \leq x}} f(u)g(v) - \sum_{\substack{u \leq M \\ v \leq x/M}} f(u)g(v),$$

yielding the required result. ■

Visually, the above argument can be understood as follows.



15.6 Average order of $d(n)$

Here we give an instance of how the Dirichlet hyperbola method provides better estimates. We start by mimicking the proof of Theorem 15.4 to estimate the partial sum of $d(n)$, the divisor function:

$$\sum_{n \leq x} d(n) = \sum_{n \leq x} \sum_{m \mid n} 1 = \sum_{\substack{m,d \\ md \leq x}} 1 = \sum_{d \leq x} \left\lfloor \frac{x}{d} \right\rfloor = \sum_{d \leq x} \left(\frac{x}{d} + O(1) \right) = x \log x + O(x).$$

Then in the next theorem, we will see that with the Dirichlet hyperbola method, the above $O(x)$ term can be explicitly expressed, and the error can be reduced to $O(\sqrt{x})$.

Theorem 15.8 As $x \to \infty$,

$$\sum_{n \leq x} d(n) = x \log x + (2\gamma - 1)x + O(\sqrt{x}). \tag{15.7}$$

Proof. Recalling the definition of $d(n)$, we have $d = \mathbf{1} * \mathbf{1}$. Now, in Theorem 15.7, we take $f = g = \mathbf{1}$, and note that $F(x) = G(x) = \lfloor x \rfloor$. Choosing $M = \sqrt{x}$ gives

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

as required. ■

16. Dirichlet series

16.1 Dirichlet series

In 1837, Lejeune Dirichlet (*Abhandlungen der Königlichen Preußischen Akademie der Wissenschaften zu Berlin* **48** (1837), 45–71) proved the following important result, which fully extends Theorems 1.2, 1.3 and 1.4.

Dirichlet's Theorem on Primes in Arithmetic Progressions There are infinitely many primes congruent to a modulo N provided that $(a,N)=1$.

To establish this result, many influential techniques in analytic number theory were introduced, one of which is the *Dirichlet series*, an infinite series associated with an arithmetic function.

Definition 16.1 Let f be an arithmetic function. The *Dirichlet series* for f is defined by

$$\sum_{n\geq 1}\frac{f(n)}{n^s},$$

where s is a complex variable.

R Following the German mathematician Bernhard Riemann, we always write complex variables s as

$$s=\sigma+it,$$

where σ and t are real. We usually call the set of complex numbers $\{s:\sigma>\sigma_0\}$ with σ_0 a given real number a *half-plane*.

As we are working on infinite series, an exigent issue is the analysis of convergence. One basic fact that will be frequently used is $|n^s|=n^\sigma$ for all positive integers n since $n^s=e^{s\log n}=n^\sigma e^{it\log n}$.

Rule 16.1 (Abscissa of Absolute Convergence). Suppose the series $\sum_{n\geq 1}|f(n)n^{-s}|$ does not converge for all s or diverge for all s. Then there exists a real number σ_a, called the *abscissa of absolute convergence*, such that the series $\sum_{n\geq 1}f(n)n^{-s}$ converges absolutely if $\sigma>\sigma_a$, but does not converge absolutely if $\sigma<\sigma_a$.

Proof. This is a direct consequence of the comparison test. ■

Lemma 16.2 Suppose that the series $\sum_{n\geq 1} f(n)n^{-s}$ converges for $s_0 = \sigma_0 + it_0$. Then this series converges for all s with $\sigma > \sigma_0$. Moreover, the convergence is uniform in every compact region contained in the half-plane $\sigma > \sigma_0$.

Proof. For convenience, we define for $1 \leq a < b$,

$$S(a,b) := \sum_{a<n\leq b} \frac{f(n)}{n^s}.$$

Since $\sum_{n\geq 1} f(n)n^{-s_0}$ converges, there exists a constant M such that the partial sum $S(x) := \sum_{n\leq x} f(n)n^{-s_0}$ satisfies $|S(x)| \leq M$ for all $x \geq 1$. By Abel's summation formula (15.1),

$$S(a,b) = \sum_{a<n\leq b} \frac{f(n)}{n^{s_0}} \frac{1}{n^{s-s_0}} = \frac{S(b)}{b^{s-s_0}} - \frac{S(a)}{a^{s-s_0}} + (s-s_0)\int_a^b \frac{S(x)}{x^{s-s_0+1}}dx.$$

Hence,

$$\begin{aligned}|S(a,b)| &\leq \frac{M}{b^{\sigma-\sigma_0}} + \frac{M}{a^{\sigma-\sigma_0}} + |s-s_0|M\int_a^b \frac{dx}{x^{\sigma-\sigma_0+1}}\\ &= \frac{M}{b^{\sigma-\sigma_0}} + \frac{M}{a^{\sigma-\sigma_0}} + \frac{M|s-s_0|}{\sigma-\sigma_0}\left(\frac{1}{a^{\sigma-\sigma_0}} - \frac{1}{b^{\sigma-\sigma_0}}\right)\\ &\leq 2Ma^{\sigma_0-\sigma}\left(1+\frac{|s-s_0|}{\sigma-\sigma_0}\right) = C\cdot a^{\sigma_0-\sigma}.\end{aligned}$$

Here, the factor $C = C(s,s_0) := 2M\left(1+\frac{|s-s_0|}{\sigma-\sigma_0}\right)$ is independent of a. Noting that $\sigma > \sigma_0$ and hence that $a^{\sigma_0-\sigma} \to 0$ as $a \to +\infty$, it follows by Cauchy's criterion that $\sum_{n\geq 1} f(n)n^{-s}$ converges for all s with $\sigma > \sigma_0$.

Further, in any compact region K contained in the half-plane $\sigma > \sigma_0$, we find that both $\sigma - \sigma_0 > 0$ and $|s-s_0|$ are bounded below and above. Hence, C can be chosen so that it only depends on K, thereby implying the uniform convergence in K. ∎

Rule 16.3 (Abscissa of Convergence). Suppose the series $\sum_{n\geq 1} f(n)n^{-s}$ does not converge for all s or diverge for all s. Then there exists a real number σ_c, called the *abscissa of convergence*, such that this series converges if $\sigma > \sigma_c$, and diverges if $\sigma < \sigma_c$.

Proof. This is a direct consequence of the first part in Lemma 16.2. ∎

Corollary 16.4 For any Dirichlet series $\sum_{n\geq 1} f(n)n^{-s}$ with σ_c finite, we have $0 \leq \sigma_a - \sigma_c \leq 1$.

Proof. It is sufficient to show that if $\sum_{n\geq 1} f(n)n^{-s}$ converges at some s_0, then it is absolutely convergent for all s with $\sigma > \sigma_0 + 1$. Noting that from the above assumption, $|f(n)n^{-s_0}|$ is bounded. Further, $|f(n)n^{-s}| = |f(n)n^{-s_0}| \cdot n^{\sigma_0-\sigma}$. Therefore, we obtain the absolute convergence by comparison with the series $\sum_{n\geq 1} n^{\sigma_0-\sigma}$. ∎

R The equality in $0 \leq \sigma_a - \sigma_c \leq 1$ can occur in both cases: **(i).** For the Riemann zeta function $\sum_{n\geq 1} \frac{1}{n^s}$, we have $\sigma_a = \sigma_c = 1$; **(ii).** For the alternating series $\sum_{n\geq 1} \frac{(-1)^n}{n^s}$, we have $\sigma_a = 1$ and $\sigma_c = 0$.

Rule 16.5 (Analyticity Theorem). Any Dirichlet series $F(s) = \sum_{n\ge 1} f(n)n^{-s}$ is analytic in its half-plane of convergence $\sigma > \sigma_c$, and its derivative $F'(s)$ is represented in this half-plane by the Dirichlet series

$$F'(s) = -\sum_{n\ge 1} \frac{f(n)\log n}{n^s}. \tag{16.1}$$

In particular, $F(s)$ and $F'(s)$ have the same abscissa of convergence and the same abscissa of absolute convergence.

Proof. Let us write $F_N(s) = \sum_{n\le N} f(n)n^{-s}$ for N positive integers. Note that $F_N(s)$ is entire since each $f(n)n^{-s}$ is entire. Also, we know from the second part in Lemma 16.2 that as $N \to \infty$, $F_N(s)$ converges to $F(s)$, uniformly in every compact region contained in the half-plane $\sigma > \sigma_0$ for any $\sigma_0 > \sigma_c$. Since σ_0 can be taken arbitrarily close to σ_c, we may also replace σ_0 by σ_c in the above conclusion. Further, such a compact convergence implies the locally uniform convergence of $F_N(s) \to F(s)$ in the open half-plane $\sigma > \sigma_c$. Karl Weierstrass's theorem on uniformly convergent sequences of analytic functions (see, for instance, E. Freitagand R. Busam,*Complex Analysis*, 2nd Edition, Theorem III.1.3, p. 106) then asserts that $F(s)$ is analytic in the half-plane $\sigma > \sigma_c$. Further, its derivative is obtained by differentiating term by term. ■

Rule 16.6 (Uniqueness Theorem). Given two Dirichlet series

$$F(s) = \sum_{n\ge 1} \frac{f(n)}{n^s} \qquad \text{and} \qquad G(s) = \sum_{n\ge 1} \frac{g(n)}{n^s},$$

both convergent for $\sigma > \sigma_0$. If $F(s) = G(s)$ for each s in an infinite sequence $\{s_k\}_{k\ge 1}$ such that $\sigma_k \to +\infty$ as $k \to \infty$, then $f(n) = g(n)$ for every n.

Proof. Note that the Dirichlet series for $h(n) = f(n) - g(n)$, denoted by $H(s)$, is also convergent for $\sigma > \sigma_0$. Meanwhile, $H(s) = F(s) - G(s)$. By Corollary 16.4, all three series are absolute convergent for $\sigma > \sigma_0 + 1$. Without loss of generality, we assume that the sequence $\{s_k\}_{k\ge 1}$ is such that $\sigma_0 + 1 < \sigma_1 < \sigma_2 < \cdots$. Supposing that $h(n)$ is not identical to zero for all n, there exists a minimal N with $h(N) \ne 0$ and $h(n) = 0$ for $n = 1, \ldots, N-1$. Noting that $H(s_k) = 0$, we have $h(N)N^{-s_k} = -\sum_{n\ge N+1} h(n)n^{-s_k}$. Hence,

$$|h(N)| \le \sum_{n\ge N+1} |h(n)| \frac{N^{\sigma_k}}{n^{\sigma_k}} = \sum_{n\ge N+1} |h(n)| \frac{N^{\sigma_1}}{n^{\sigma_1}} \left(\frac{N}{n}\right)^{\sigma_k - \sigma_1} \le \left(\sum_{n\ge N+1} |h(n)| \frac{N^{\sigma_1}}{n^{\sigma_1}}\right) \left(\frac{N}{N+1}\right)^{\sigma_k - \sigma_1}.$$

Note that $\sum_{n\ge N+1} |h(n)| \frac{N^{\sigma_1}}{n^{\sigma_1}}$ is a finite constant, independent of k. Letting $k \to \infty$ so that $(\sigma_k - \sigma_1) \to \infty$, we have $\left(\frac{N}{N+1}\right)^{\sigma_k - \sigma_1} \to 0$ and hence $h(N) = 0$. This leads to a contradiction, thereby implying that $h(n) = 0$, i.e. $f(n) = g(n)$, for all n. ■

16.2 Multiplication of Dirichlet series

Definition 16.2 For any arithmetic function f, we denote by $D(f;s)$ the Dirichlet series for f, namely,

$$D(f;s) := \sum_{n\ge 1} \frac{f(n)}{n^s}.$$

Theorem 16.7 Let f and g be arithmetic functions such that $D(f;s)$ and $D(g;s)$ have finite abscissas of absolute convergence. In the half-plane where both $D(f;s)$ and $D(g;s)$ converge absolutely, we have that $D(f*g;s)$ also converges absolutely in this half-plane, and that

$$D(f;s)D(g;s) = D(f*g;s), \tag{16.2}$$

where $f*g$ is the Dirichlet convolution of f and g.

Proof. Since the series $D(f;s)$ and $D(g;s)$ are absolutely convergent in the half-plane, so is their Cauchy product, which has the same value as $D(f;s)D(g;s)$. Note that the Cauchy product of $D(f;s) = \sum_{m\geq 1}\frac{f(m)}{m^s}$ and $D(g;s) = \sum_{n\geq 1}\frac{g(n)}{n^s}$ equals

$$\sum_{k\geq 2}\sum_{\substack{m,n\geq 1\\ m+n=k}}\frac{f(m)g(n)}{(mn)^s} = \sum_{\ell\geq 1}\sum_{\substack{m,n\geq 1\\ mn=\ell}}\frac{f(m)g(n)}{(mn)^s} = \sum_{\ell\geq 1}\frac{(f*g)(\ell)}{\ell^s} = D(f*g;s),$$

in which the first equality is valid as the absolute convergence allows us to rearrange the terms without altering the sum. The desired result therefore follows. ■

Corollary 16.8 Let f be an arithmetic function with $f(1)\neq 0$, and let f^{-1} be the Dirichlet inverse of f. Then in any half-plane where $D(f;s)$ and $D(f^{-1};s)$ converge absolutely, we have $D(f;s)\neq 0$ and $D(f^{-1};s)\neq 0$. Also,

$$D(f^{-1};s) = \frac{1}{D(f;s)}. \tag{16.3}$$

Proof. We use the fact that $f*f^{-1} = \varepsilon$. Hence, by Theorem 16.7, $D(f;s)D(f^{-1};s) = D(\varepsilon;s) = 1$. ■

16.3 Dirichlet series for some arithmetic functions

Now we present some examples of the Dirichlet series.

■ **Example 16.1** The Dirichlet series for the constant function $\mathbf{1}$ is the *Riemann zeta function*

$$D(\mathbf{1};s) = \sum_{n\geq 1}\frac{1}{n^s} = \zeta(s). \tag{16.4}$$

Meanwhile, $D(\mathbf{1};s)$ has abscissa of absolute convergence $\sigma_a = 1$ and abscissa of convergence $\sigma_c = 1$. ■

■ **Example 16.2** The Dirichlet series for the unit function ε is

$$D(\varepsilon;s) = 1. \tag{16.5}$$

Meanwhile, $D(\varepsilon;s)$ is absolutely convergent in $\mathbb{C}$. ■

■ **Example 16.3** The Dirichlet series for the identity function **id** is

$$D(\mathrm{id};s) = \sum_{n\geq 1}\frac{1}{n^{s-1}} = \zeta(s-1). \tag{16.6}$$

Meanwhile, $D(\mathrm{id};s)$ has abscissa of absolute convergence $\sigma_a = 2$ and abscissa of convergence $\sigma_c = 2$. Further, if we define $\mathrm{id}^\alpha(n) = n^\alpha$ for all n with $\alpha \in \mathbb{C}$, then the Dirichlet series for id^α is

$$D(\mathrm{id}^\alpha;s) = \sum_{n\geq 1} \frac{1}{n^{s-\alpha}} = \zeta(s-\alpha). \tag{16.7}$$

Meanwhile, $D(\mathrm{id}^\alpha;s)$ has abscissa of absolute convergence $\sigma_a = 1+\Re(\alpha)$ and abscissa of convergence $\sigma_c = 1+\Re(\alpha)$. ■

■ **Example 16.4** The Dirichlet series for the Möbius function μ is

$$D(\mu;s) = \sum_{n\geq 1} \frac{\mu(n)}{n^s} = \frac{1}{\zeta(s)}, \tag{16.8}$$

for $\sigma > 1$. This is because μ is the Dirichlet inverse of $\mathbf{1}$, i.e. $\mu * \mathbf{1} = \varepsilon$. ■

■ **Example 16.5** The Dirichlet series for the divisor function σ_α is

$$D(\sigma_\alpha;s) = \sum_{n\geq 1} \frac{\sigma_\alpha(n)}{n^s} = \zeta(s)\zeta(s-\alpha), \tag{16.9}$$

for $\sigma > \max\{1, 1+\Re(\alpha)\}$. This is because $\sigma_\alpha = \mathbf{1} * \mathrm{id}^\alpha$, and hence $D(\sigma_\alpha;s) = D(\mathbf{1};s)D(\mathrm{id}^\alpha;s)$. In particular, the Dirichlet series for the number-of-divisors function d is

$$D(d;s) = \sum_{n\geq 1} \frac{d(n)}{n^s} = \zeta(s)^2, \tag{16.10}$$

for $\sigma > 1$, and the Dirichlet series for the sum-of-divisors function σ is

$$D(\sigma;s) = \sum_{n\geq 1} \frac{\sigma(n)}{n^s} = \zeta(s)\zeta(s-1), \tag{16.11}$$

for $\sigma > 2$. ■

■ **Example 16.6** The Dirichlet series for Euler's totient function ϕ is

$$D(\phi;s) = \sum_{n\geq 1} \frac{\phi(n)}{n^s} = \frac{\zeta(s-1)}{\zeta(s)}, \tag{16.12}$$

for $\sigma > 2$. This is because $\phi = \mu * \mathrm{id}$ by (13.5), and hence $D(\phi;s) = D(\mu;s)D(\mathrm{id};s)$. ■

■ **Example 16.7** The Dirichlet series for the logarithm function log is

$$D(\log;s) = \sum_{n\geq 1} \frac{\log(n)}{n^s} = -\zeta'(s), \tag{16.13}$$

where we make use of (16.1). Meanwhile, $D(\log;s)$ has abscissa of absolute convergence $\sigma_a = 1$ and abscissa of convergence $\sigma_c = 1$. ■

■ **Example 16.8** The Dirichlet series for the Mangoldt function Λ is

$$D(\Lambda;s) = \sum_{n\geq 1} \frac{\Lambda(n)}{n^s} = -\frac{\zeta'(s)}{\zeta(s)}, \tag{16.14}$$

for $\sigma > 1$. This is because $\Lambda = \mu * \log$ by (13.8), and hence $D(\Lambda;s) = D(\mu;s)D(\log;s)$. ■

16.4 Euler products

Recall that in our proof of the divergence of $\sum_p \frac{1}{p}$ in Sect. 1.6, we make use of the following relation

$$\prod_{p\le N}\left(1+\frac{1}{p}+\frac{1}{p^2}+\cdots\right)=\sum_{\substack{n\ge 1\\ n\text{ has no prime factor }>N}}\frac{1}{n}. \tag{16.15}$$

This idea was first discovered by Euler, and in 1737 he proved the following theorem, also known as the *analytic version of the fundamental theorem of arithmetic*.

Theorem 16.9 (Euler). Let a be a multiplicative function such that the series $\sum_{n\ge1} a(n)$ is absolutely convergent. Then this series can be expressed as an absolutely convergent infinite product indexed by prime numbers,

$$\sum_{n\ge1} a(n)=\prod_p\big(1+a(p)+a(p^2)+\cdots\big). \tag{16.16}$$

In particular, if a is completely multiplicative, we have

$$\sum_{n\ge1} a(n)=\prod_p \frac{1}{1-a(p)}. \tag{16.17}$$

R The infinite product in (16.16) is called the *Euler product* of the series $\sum_{n\ge1} a(n)$.

Proof. We elaborate our argument for (16.15) that was originally presented in Sect. 1.6. Recall always that $a(1)=1$ since a is multiplicative. Let us define the partial product

$$P(N):=\prod_{p\le N}\big(1+a(p)+a(p^2)+\cdots\big).$$

Note that for each p, the series $\sum_{k\ge0} a(p^k)$ is absolutely convergent as $\sum_{n\ge1} a(n)$ converges absolutely. As a consequence, we may expand the product and rearrange the terms. On the other hand, for any n with the canonical form $n=\prod_j p_j^{\alpha_j}$ such that no prime factor is greater than N, i.e. $p_j\le N$ for all j, we find that $a(n)=\prod_j a(p_j^{\alpha_j})$ since a is multiplicative, and that it corresponds to exactly one term in the expansion of $P(N)$. Thus,

$$P(N)=\sum_{\substack{n\ge1\\ n\text{ has no prime factor }>N}} a(n),$$

or equivalently,

$$\sum_{n\ge1} a(n)-P(N)=\sum_{\substack{n\ge1\\ n\text{ has at least one prime factor }>N}} a(n).$$

Hence, recalling that $\sum_{n\ge1}|a(n)|$ converges, we have

$$\left|\sum_{n\ge1} a(n)-P(N)\right|\le\sum_{n>N}|a(n)|\to 0 \qquad (\text{as } N\to\infty),$$

thereby implying that $P(N)\to\sum_{n\ge1} a(n)$ as $N\to\infty$.

Now let us show that the infinite product in (16.16) is absolutely convergent. To see this, it is sufficient to prove that $\sum_p |u_p|$ converges where $u_p = a(p) + a(p^2) + \cdots$. This is obvious since

$$\sum_p |u_p| \leq \sum_p (|a(p)| + |a(p^2)| + \cdots) \leq \sum_{n\geq 2} |a(n)|,$$

while $\sum_{n\geq 2} |a(n)|$ is finite by the absolute convergence of $\sum_{n\geq 1} a(n)$.

Finally, when a is completely multiplicative, we have $a(p^k) = a(p)^k$ for all prime powers p^k. Therefore, the absolutely convergent subseries $\sum_{k\geq 0} a(p^k) = \sum_{k\geq 0} a(p)^k$ can be evaluated as a geometric series and hence equals $\frac{1}{1-a(p)}$. ■

Corollary 16.10 Let $\sum_{n\geq 1} f(n)n^{-s}$ be a Dirichlet series that converges absolutely in the half-plane $\sigma > \sigma_a$. If f is multiplicative, then for $\sigma > \sigma_a$,

$$\sum_{n\geq 1} \frac{f(n)}{n^s} = \prod_p \left(1 + \frac{f(p)}{p^s} + \frac{f(p^2)}{p^{2s}} + \cdots\right). \tag{16.18}$$

In particular, if f is completely multiplicative, then for $\sigma > \sigma_a$,

$$\sum_{n\geq 1} \frac{f(n)}{n^s} = \prod_p \frac{1}{1 - f(p)p^{-s}}. \tag{16.19}$$

Proof. We simply use the fact that if $f(n)$ is multiplicative or completely multiplicative, so is $f(n)n^{-s}$. ■

■ **Example 16.9** We have the following Euler product expressions:

(i) $\displaystyle\sum_{n\geq 1} \frac{1}{n^s} = \zeta(s) = \prod_p \frac{1}{1-p^{-s}}$ for $\sigma > 1$;

(ii) $\displaystyle\sum_{n\geq 1} \frac{\mu(n)}{n^s} = \frac{1}{\zeta(s)} = \prod_p (1-p^{-s})$ for $\sigma > 1$;

(iii) $\displaystyle\sum_{n\geq 1} \frac{\sigma_\alpha(n)}{n^s} = \zeta(s)\zeta(s-\alpha) = \prod_p \frac{1}{(1-p^{-s})(1-p^{\alpha-s})}$ for $\sigma > \max\{1, 1+\Re(\alpha)\}$;

(iv) $\displaystyle\sum_{n\geq 1} \frac{d(n)}{n^s} = \zeta(s)^2 = \prod_p \frac{1}{(1-p^{-s})^2}$ for $\sigma > 1$;

(v) $\displaystyle\sum_{n\geq 1} \frac{\sigma(n)}{n^s} = \zeta(s)\zeta(s-1) = \prod_p \frac{1}{(1-p^{-s})(1-p^{1-s})}$ for $\sigma > 2$;

(vi) $\displaystyle\sum_{n\geq 1} \frac{\phi(n)}{n^s} = \frac{\zeta(s-1)}{\zeta(s)} = \prod_p \frac{1-p^{-s}}{1-p^{1-s}}$ for $\sigma > 2$.

■

17. Dirichlet characters

17.1 Dirichlet characters

For the purpose of proving Dirichlet's theorem on primes in arithmetic progressions, another crucial tool is the *Dirichlet character*.

Definition 17.1 Let N be a positive integer. A *Dirichlet character* or a *character modulo* N is a complex-valued arithmetic function $\chi : \mathbb{Z}_{>0} \to \mathbb{C}$ with the following properties:

(i) $\chi(ab) = \chi(a)\chi(b)$, i.e. χ is completely multiplicative;

(ii) $\chi(a) \begin{cases} = 0 & \text{if } (a,N) > 1, \\ \neq 0 & \text{if } (a,N) = 1; \end{cases}$

(iii) $\chi(a+N) = \chi(a)$, i.e. χ is periodic with period N.

R We sometimes define Dirichlet characters on $\mathbb{Z}$ instead of $\mathbb{Z}_{>0}$ by the same conditions.

Example 17.1 For each positive integer N,

$$\chi_0(a) = \chi_{N,0}(a) = \begin{cases} 0 & \text{if } (a,N) > 1 \\ 1 & \text{if } (a,N) = 1 \end{cases}$$

is a Dirichlet character modulo N. We call this character the *principal character*. We shall label Dirichlet characters modulo N by $\chi_{N,0}$, $\chi_{N,1}$, ..., or by χ_0, χ_1, ... if there is no ambiguity concerning the modulus.

Theorem 17.1 Let N be a positive integer and a be such that $(a,N) = 1$. Then for any character χ modulo N, $\chi(a)$ is a $\phi(N)$-th root of unity.

Proof. By the Fermat–Euler theorem, $a^{\phi(N)} \equiv 1 \pmod{N}$. Hence, $\chi(a)^{\phi(N)} = \chi(a^{\phi(N)}) = \chi(1) = 1$, where we use the fact that χ is completely multiplicative. ■

From Theorem 17.1, we see that if $\chi(a)$ is a real number, then it takes value only from $\{-1, 0, 1\}$.

Definition 17.2 Let χ be a Dirichlet character modulo a positive integer N. If all of its values are real, we say χ is a *real character*. Otherwise, it is called a *complex character*.

Example 17.2 The principal character modulo any N is real. Another example of real characters is the Legendre symbol $\left(\frac{a}{p}\right)$ where the modulus $N = p$ is an odd prime. We will give instances of complex characters in later sections. ■

Theorem 17.2 Let N be a positive integer.

(i) If χ and χ' are two characters modulo N, so is their product $\chi\chi'$, defined by $\chi\chi'(a) := \chi(a)\chi'(a)$.

(ii) If χ is a character modulo N, so is its complex conjugate $\overline{\chi}$, defined by $\overline{\chi}(a) := \overline{\chi(a)}$, the complex conjugate of $\chi(a)$. In particular, $\chi\overline{\chi} = \chi_0$, the principal character.

Proof. The results follow by a direct verification of the three conditions in Definition 17.1. For the second part in (ii), we also use the fact that $z\bar{z} = |z|^2$ for any complex z, and any root of unity has absolute value 1. ■

Corollary 17.3 Let N be a positive integer. If χ is a real character modulo N, then $\chi^2 = \chi_0$, the principal character.

Proof. This is because for real χ, we have $\chi(a) \in \{-1, 1\}$ for all $(a, N) = 1$. Hence, $\chi^2(a) = \chi(a)^2 = (\pm 1)^2 = 1 = \chi_0(a)$. ■

Theorem 17.4 Let N be a positive integer and a be such that $(a, N) = 1$. Let $\bar{a}$ be the inverse of a modulo N, i.e. $a\bar{a} \equiv 1 \pmod{N}$. Then for any character χ modulo N, $\chi(\bar{a}) = \chi(a)^{-1} = \overline{\chi(a)}$.

Proof. Noting that $a\bar{a} \equiv 1 \pmod{N}$, we have $\chi(a)\chi(\bar{a}) = \chi(a\bar{a}) = \chi(1) = 1$. Hence, $\chi(\bar{a}) = \chi(a)^{-1}$. Also, for any complex z with $|z| = 1$, we have $z^{-1} = \bar{z}$, giving the second equality. ■

17.2 Construction of Dirichlet characters modulo prime powers

Definition 17.3 For positive integers n, we define $\zeta_n := e^{\frac{2\pi i}{n}}$.

Construction 17.1 Let $N = 2$, or 4, or p^α with p an odd prime and α a positive integer. Let g be a primitive root of N. We know that $\{g^0, g^1, \ldots, g^{\phi(N)-1}\}$ gives a reduced system modulo N. For each a with $(a, N) = 1$, we may find a unique integer d with $0 \leq d < \phi(N)$ such that $a \equiv g^d \pmod{N}$. We call this d the *index of a modulo N with respect to g*, denoted by $\operatorname{ind} a = \operatorname{ind}_{N:g} a = d$. For any character χ modulo N, we know from Theorem 17.1 that $\chi(g)$ is a $\phi(N)$-th root of unity. We claim that this character χ is uniquely determined by $\chi(g)$. This is because for any a with $(a, N) = 1$, we have $\chi(a) = \chi(g^{\operatorname{ind} a}) = \chi(g)^{\operatorname{ind} a}$. ■

Example 17.3 For $N = 2$, we choose the primitive root $g = 1$; for $N = 3$, we choose the primitive root $g = 2$; for $N = 5$, we choose the primitive root $g = 2$. ■

a	1
$\operatorname{ind}_{2:1} a$	0

χ \ a	1
$\chi_{2,0}$	1

a	1	2
$\operatorname{ind}_{3:2} a$	0	1

χ \ a	1	2
$\chi_{3,0}$	1	1
$\chi_{3,1}$	1	-1

a	1	2	3	4
$\operatorname{ind}_{5:2} a$	0	1	3	2

χ \ a	1	2	3	4
$\chi_{5,0}$	1	1	1	1
$\chi_{5,1}$	1	i	$-i$	-1
$\chi_{5,2}$	1	-1	-1	1
$\chi_{5,3}$	1	$-i$	i	-1

For $N = 2^\alpha$ with $\alpha \geq 3$, however, we know from Theorem 5.16 that N has no primitive roots. Hence, a different construction is necessary.

Lemma 17.5 For $\alpha \geq 3$, we have $\mathrm{ord}_{2^\alpha} 5 = 2^{\alpha-2}$.

Proof. We have seen from the proof of Theorem 5.16 that $5^{2^{\alpha-2}} \equiv 1 \pmod{2^\alpha}$. Now, it suffices to show that $5^{2^{\alpha-3}} = 1 + 2^{\alpha-1}x$ with $2 \nmid x$, so that $5^{2^{\alpha-3}} \not\equiv 1 \pmod{2^\alpha}$. We prove this claim by induction on α. For $\alpha = 3$, we have $5^{2^0} = 5 = 1 + 2^2 \cdot 1$. Now we assume that the claim is true for some $\alpha \geq 3$, and we prove the $\alpha + 1$ case. Note that

$$5^{2^{(\alpha+1)-3}} = \left(5^{2^{\alpha-3}}\right)^2 = \left(1 + 2^{\alpha-1}x\right)^2 = 1 + 2^\alpha\left(x + 2^{\alpha-2}x^2\right).$$

Here, $x + 2^{\alpha-2}x^2$ is odd since x is odd and $\alpha \geq 3$. We remark that the above argument also gives another confirmation of $5^{2^{\alpha-2}} \equiv 1 \pmod{2^\alpha}$. ■

Lemma 17.6 Let $N = 2^\alpha$ with $\alpha \geq 3$. For every odd integer a, there exist unique integers $\nu_{N:-1}(a)$ and $\nu_{N:5}(a)$ with $0 \leq \nu_{N:-1}(a) < 2$ and $0 \leq \nu_{N:5}(a) < 2^{\alpha-2}$ such that $a \equiv (-1)^{\nu_{N:-1}(a)} 5^{\nu_{N:5}(a)} \pmod{N}$.

Proof. It suffices to show that $\{(-1)^u 5^v : 0 \leq u < 2 \text{ and } 0 \leq v < 2^{\alpha-2}\}$ is a reduced system modulo $N = 2^\alpha$. First, the $2^{\alpha-2}$ numbers 5^v (with $0 \leq v < 2^{\alpha-2}$) are pairwise incongruent modulo N since $\mathrm{ord}_{2^\alpha} 5 = 2^{\alpha-2}$ by Lemma 17.5. The same property also holds for the $2^{\alpha-2}$ numbers -5^v (with $0 \leq v < 2^{\alpha-2}$). Finally, we see that $5^{u_1} \not\equiv -5^{u_2} \pmod{N}$ since $5^{u_1} \equiv 1 \pmod{4}$, while $-5^{u_1} \equiv 3 \pmod{4}$, where we recall that $4 \mid N$. ■

■ **Construction 17.2** Let $N = 2^\alpha$ with $\alpha \geq 3$. For any character χ modulo N, we find that $\chi(-1)^2 = \chi((-1)^2) = \chi(1) = 1$, implying that $\chi(-1)$ is a quadratic root of unity. Also, since $\mathrm{ord}_N 5 = 2^{\alpha-2} = \frac{\phi(N)}{2}$ by Lemma 17.5, we have that $\chi(5)$ is a $\frac{\phi(N)}{2}$-th root of unity. We claim that this character χ is uniquely determined by $\chi(-1)$ and $\chi(5)$. This is because for any a with $(a,N) = 1$, we know from Lemma 17.6 that $\chi(a) = \chi((-1)^{\nu_{N:-1}(a)} 5^{\nu_{N:5}(a)}) = \chi(-1)^{\nu_{N:-1}(a)} \chi(5)^{\nu_{N:5}(a)}$. ■

■ **Example 17.4** For $N = 2^3 = 8$, we have $\chi(-1) \in \{1, -1\}$ and $\chi(5) \in \{1, -1\}$. We write the characters modulo N as $\chi_{(\chi(-1);\chi(5))}$ for clarity. ■

a	1	3	5	7
$\nu_{8:-1}(a)$	0	1	0	1
$\nu_{8:5}(a)$	0	1	1	0

a χ	1	3	5	7
$\chi_{(1;1)}$	1	1	1	1
$\chi_{(1;-1)}$	1	-1	-1	1
$\chi_{(-1;1)}$	1	-1	1	-1
$\chi_{(-1;-1)}$	1	1	-1	-1

Corollary 17.7 Let N be a prime power. Then there are exactly $\phi(N)$ characters modulo N. In particular, for any a with $(a,N) = 1$ and $a \not\equiv 1 \pmod{N}$, there always exists a character χ such that $\chi(a) \neq 1$.

Proof. For $N = 2$, or 4, or p^α with p an odd prime and α a positive integer, the first part comes from the fact that the number of $\phi(N)$-th roots of unity is $\phi(N)$, namely, $\zeta_{\phi(N)}^0 = 1$, $\zeta_{\phi(N)}^1$, $\ldots$, $\zeta_{\phi(N)}^{\phi(N)-1}$. Hence, there are exactly $\phi(N)$ choices of $\chi(g)$ as in Construction 17.1, and thus exactly $\phi(N)$ characters modulo N. Finally, for any a with $(a,N) = 1$ and $a \not\equiv 1 \pmod{N}$, we know that $0 < \mathrm{ind}\, a < \phi(N)$. Hence, we choose a character χ such that $\chi(g) = \zeta_{\phi(N)}$, and thus $\chi(a) = \zeta_{\phi(N)}^{\mathrm{ind}\, a} \neq 1$.

For $N = 2^\alpha$ with $\alpha \geq 3$, the first part comes from the fact that the number of quadratic roots of unity is 2, namely 1 and -1; and the number of $\frac{\phi(N)}{2}$-th roots of unity is $\frac{\phi(N)}{2}$, namely, $\zeta_{\phi(N)/2}^{0}=1$, $\zeta_{\phi(N)/2}^{1}$, $\ldots$, $\zeta_{\phi(N)/2}^{\phi(N)/2-1}$. Hence, there are exactly 2 choices of $\chi(-1)$ and exactly $\frac{\phi(N)}{2}$ choices of $\chi(5)$ as in Construction 17.2, and thus exactly $2\cdot\frac{\phi(N)}{2}=\phi(N)$ characters modulo N. Finally, for any a with $(a,N)=1$ and $a\not\equiv 1 \pmod N$, we see from Lemma 17.6 that at least one of $\nu_{N:-1}(a)$ and $\nu_{N:5}(a)$ is not zero. If $\nu_{N:-1}(a)\neq 0$ (and hence $\nu_{N:-1}(a)=1$), we choose a character χ such that $\chi(-1)=-1$ and $\chi(5)=1$, and thus $\chi(a)=(-1)^1\cdot 1=-1\neq 1$; if $\nu_{N:5}(a)\neq 0$ (and hence $0<\nu_{N:5}(a)<\frac{\phi(N)}{2}$), we choose a character χ such that $\chi(-1)=1$ and $\chi(5)=\zeta_{\phi(N)/2}$, and thus $\chi(a)=1\cdot\zeta_{\phi(N)/2}^{\nu_{N:5}(a)}\neq 1$. ■

17.3 Construction of Dirichlet characters modulo generic integers

Now we construct characters modulo generic integers.

Lemma 17.8 Let m and n be positive integers such that $(m,n)=1$, and write $N=mn$. There exist a unique reduced system $R_n(m):=\{r_{m,1},\ldots,r_{m,\phi(m)}\}$ modulo m such that $1\leq r_{m,i}\leq N$ and $r_{m,i}\equiv 1 \pmod n$ for all i, and a unique reduced system $R_m(n):=\{r_{n,1},\ldots,r_{n,\phi(n)}\}$ modulo n such that $1\leq r_{n,j}\leq N$ and $r_{n,j}\equiv 1 \pmod m$ for all j. In particular,

(i) $(r_{m,i},N)=1$ for all i and $(r_{n,j},N)=1$ for all j;

(ii) $R_n(m)\cap R_m(n)=\{1\}$.

Proof. By the Chinese Remainder Theorem, the system

$$\begin{cases} x_a\equiv a & \pmod m\\ x_a\equiv 1 & \pmod n\end{cases}$$

has a unique solution modulo N. Running a over a reduced system modulo m and choosing the solutions x_a so that $1\leq x_a\leq N$, we arrive at the unique reduced system $R_n(m)$ modulo m. Similarly, we have the unique reduced system $R_m(n)$ modulo n. Further, $(r_{m,i},m)=1$ by definition. Also, $r_{m,i}\equiv 1 \pmod n$ implies that $(r_{m,i},n)=1$. Hence, $(r_{m,i},N)=1$. By symmetry, we also have $(r_{n,j},N)=1$. Finally, if $r\in R_n(m)\cap R_m(n)$, then $r\equiv 1 \pmod m$ and $r\equiv 1 \pmod n$, and hence the only possibility is $r=1$. ■

Lemma 17.9 Let m and n be positive integers such that $(m,n)=1$, and write $N=mn$. Let $R_n(m)=\{r_{m,1},\ldots,r_{m,\phi(m)}\}$ and $R_m(n)=\{r_{n,1},\ldots,r_{n,\phi(n)}\}$ be as in Lemma 17.8. Then for any a such that $(a,N)=1$, there are unique integers $r_{m,i}\in R_n(m)$ and $r_{n,j}\in R_m(n)$ such that $a\equiv r_{m,i}r_{n,j} \pmod N$.

Proof. Note that there are $\phi(m)\phi(n)=\phi(N)$ such $r_{m,i}r_{n,j}$. Further, by Lemma 17.8(i), $(r_{m,i}r_{n,j},N)=1$. Now, it suffices to show that they are pairwise incongruent modulo N. If we have $r_{m,i}r_{n,j}\equiv r_{m,i'}r_{n,j'} \pmod N$, then it implies that $r_{m,i}r_{n,j}\equiv r_{m,i'}r_{n,j'} \pmod m$ and hence $r_{m,i}\equiv r_{m,i'} \pmod m$ since $r_{n,j}\equiv r_{n,j'}\equiv 1 \pmod m$. Similarly, we have $r_{n,j}\equiv r_{n,j'} \pmod n$. The desired result thus follows. ■

■ **Construction 17.3** Let m and n be positive integers such that $(m,n)=1$, and write $N=mn$. Let $R_n(m)=\{r_{m,1},\ldots,r_{m,\phi(m)}\}$ and $R_m(n)=\{r_{n,1},\ldots,r_{n,\phi(n)}\}$ be as in Lemma 17.8. For each character χ' modulo m and each character χ'' modulo n, we define $[\chi',\chi'']=:\chi$ by

$$\chi(a)=\begin{cases}0 & \text{if } (a,N)>1,\\ \chi'(r_{m,i})\chi''(r_{n,j}) & \text{if } (a,N)=1,\end{cases}$$

where we use Lemma 17.9 to write $a\equiv r_{m,i}r_{n,j} \pmod N$ for the second case. ■

Theorem 17.10 The function χ as in Construction 17.3 is a character modulo N.

Proof. It is sufficient to show that χ is completely multiplicative. In particular, given a and b with $(a,N) = (b,N) = 1$, we want to show that $\chi(ab) = \chi(a)\chi(b)$. By Lemma 17.9, we write $a \equiv r_{m,i_1} r_{n,j_1} \pmod{N}$, $b \equiv r_{m,i_2} r_{n,j_2} \pmod{N}$ and $ab \equiv r_{m,I} r_{n,J} \pmod{N}$. Hence, $r_{m,I} r_{n,J} \equiv r_{m,i_1} r_{m,i_2} r_{n,j_1} r_{n,j_2} \pmod{N}$, and further $r_{m,I} r_{n,J} \equiv r_{m,i_1} r_{m,i_2} r_{n,j_1} r_{n,j_2} \pmod{m}$. Since $r_{n,J} \equiv r_{n,j_1} \equiv r_{n,j_2} \equiv 1 \pmod{m}$, we have $r_{m,I} \equiv r_{m,i_1} r_{m,i_2} \pmod{m}$, and therefore, $\chi'(r_{m,I}) = \chi'(r_{m,i_1} r_{m,i_2}) = \chi'(r_{m,i_1})\chi'(r_{m,i_2})$. Similarly, $\chi''(r_{n,J}) = \chi''(r_{n,j_1})\chi''(r_{n,j_2})$. It follows that

$$\chi(ab) = \chi'(r_{m,I})\chi''(r_{n,J}) = \chi'(r_{m,i_1})\chi'(r_{m,i_2})\chi''(r_{n,j_1})\chi''(r_{n,j_2}) = \chi(a)\chi(b),$$

as required. ■

■ **Example 17.5** For $N = 3 \cdot 5 = 15$, we first find that $R_5(3) = \{1, 11\}$ and $R_3(5) = \{1, 7, 13, 4\}$, and then compute $r_{3,i} r_{5,j}$ mod 15 for each i and j. The characters modulo 3 and 5 are given in Example 17.3. ■

$r_{3,i}$ \ $r_{5,j}$	1	7	13	4
1	1	7	13	4
11	11	2	8	14

χ \ a	1	2	4	7	8	11	13	14
$[\chi_{3,0}, \chi_{5,0}]$	1	1	1	1	1	1	1	1
$[\chi_{3,0}, \chi_{5,1}]$	1	i	-1	i	$-i$	1	$-i$	-1
$[\chi_{3,0}, \chi_{5,2}]$	1	-1	1	-1	-1	1	-1	1
$[\chi_{3,0}, \chi_{5,3}]$	1	$-i$	-1	$-i$	i	1	i	-1
$[\chi_{3,1}, \chi_{5,0}]$	1	-1	1	1	-1	-1	1	-1
$[\chi_{3,1}, \chi_{5,1}]$	1	$-i$	-1	i	i	-1	$-i$	1
$[\chi_{3,1}, \chi_{5,2}]$	1	1	1	-1	1	-1	-1	-1
$[\chi_{3,1}, \chi_{5,3}]$	1	i	-1	$-i$	$-i$	-1	i	1

The following are implications of Construction 17.3.

Theorem 17.11 Let m and n be positive integers such that $(m,n) = 1$, and write $N = mn$. If $[\chi', \chi''] = [\hat{\chi}', \hat{\chi}'']$ with χ' and $\hat{\chi}'$ characters modulo m, and χ'' and $\hat{\chi}''$ characters modulo m, then $\chi' = \hat{\chi}'$ and $\chi'' = \hat{\chi}''$.

Proof. For each $r_{m,i} \in R_n(m)$, we note that $r_{m,i} \equiv r_{m,i} \cdot 1 \pmod{N}$ while $1 \in R_m(n)$. Hence, $[\chi', \chi''] = [\hat{\chi}', \hat{\chi}'']$ implies that $[\chi', \chi''](r_{m,i}) = [\hat{\chi}', \hat{\chi}''](r_{m,i})$, or $\chi'(r_{m,i})\chi''(1) = \hat{\chi}'(r_{m,i})\hat{\chi}''(1)$, or $\chi'(r_{m,i}) = \hat{\chi}'(r_{m,i})$. Since $R_n(m)$ is a reduced system modulo m, we have $\chi' = \hat{\chi}'$. By symmetry, we also have $\chi'' = \hat{\chi}''$. ■

Theorem 17.12 Let m and n be positive integers such that $(m,n) = 1$, and write $N = mn$. Let $\chi = [\chi', \chi'']$. If $\chi' = \chi_{m,0}$, the principal character modulo m, then for any a with $(a,N) = 1$, we have $\chi(a) = \chi''(a)$. Also, if $\chi'' = \chi_{n,0}$, the principal character modulo n, then for any a with $(a,N) = 1$, we have $\chi(a) = \chi'(a)$.

Proof. For any a with $(a,N) = 1$. We write $a \equiv r_{m,i} r_{n,j} \pmod{N}$ as in Construction 17.3, and hence $a \equiv r_{m,i} r_{n,j} \pmod{n}$. Noting that $r_{m,i} \equiv 1 \pmod{n}$ by definition, we have $a \equiv r_{n,j} \pmod{n}$. If $\chi' = \chi_{m,0}$, then $\chi(a) = \chi_{m,0}(r_{m,i})\chi''(r_{n,j}) = \chi''(r_{n,j}) = \chi''(a)$, as required. We further derive the second part by symmetry. ■

Theorem 17.13 Let m and n be positive integers such that $(m,n) = 1$, and write $N = mn$. There are no characters modulo N other than those as constructed in Construction 17.3.

Proof. To show that there are no other characters modulo N, it suffices to prove that for each character χ modulo N, we can find a character χ' modulo m and a character χ'' modulo n such that $\chi = [\chi', \chi'']$. Let $R_n(m) = \{r_{m,1}, \ldots, r_{m,\phi(m)}\}$ and $R_m(n) = \{r_{n,1}, \ldots, r_{n,\phi(n)}\}$ be as in

Construction 17.3. Recall that they form a reduced system modulo m and n, respectively. We first define

$$\chi|_m(a) = \begin{cases} 0 & \text{if } (a,m) > 1, \\ \chi(r_{m,i}) & \text{if } (a,m) = 1 \text{ and } a \equiv r_{m,i} \pmod{m}. \end{cases}$$

We claim that $\chi|_m$ is a character modulo m. In fact, it suffices to show that $\chi|_m$ is completely multiplicative. For any a and b with $(a,m) = (b,m) = 1$, we assume that $a \equiv r_{m,i'} \pmod{m}$, $b \equiv r_{m,i''} \pmod{m}$ and $ab \equiv r_{m,I} \pmod{m}$. Then $r_{m,I} \equiv r_{m,i'}r_{m,i''} \pmod{m}$. Also, $r_{m,I} \equiv 1 \equiv r_{m,i'}r_{m,i''} \pmod{n}$ by definition. Hence, by the Chinese Remainder Theorem, $r_{m,I} \equiv r_{m,i'}r_{m,i''} \pmod{N}$. We then have

$$\chi|_m(ab) = \chi(r_{m,I}) = \chi(r_{m,i'}r_{m,i''}) = \chi(r_{m,i'})\chi(r_{m,i''}) = \chi|_m(a)\chi|_m(b),$$

as required. Similarly, we define

$$\chi|_n(a) = \begin{cases} 0 & \text{if } (a,n) > 1, \\ \chi(r_{n,j}) & \text{if } (a,n) = 1 \text{ and } a \equiv r_{n,j} \pmod{n}, \end{cases}$$

and find that $\chi|_n$ is a character modulo n. Finally, we claim that $\chi = [\chi|_m, \chi|_n]$. In fact, if we write $[\chi|_m, \chi|_n] = \tilde{\chi}$, then for any a with $(a,N) = 1$ and $a \equiv r_{m,i}r_{n,j} \pmod{N}$,

$$\tilde{\chi}(a) = \chi|_m(r_{m,i})\chi|_n(r_{n,j}) = \chi(r_{m,i})\chi(r_{n,j}) = \chi(r_{m,i}r_{n,j}) = \chi(a),$$

as desired. ■

Corollary 17.14 Let m and n be positive integers such that $(m,n) = 1$, and write $N = mn$.

(i) If there are A characters modulo m and B characters modulo n, provided that A and B are finite, then there are AB characters modulo N;

(ii) If for each u with $(u,m) = 1$ and $u \not\equiv 1 \pmod{m}$ there exists a character χ' modulo m such that $\chi'(u) \neq 1$, and for each v with $(v,n) = 1$ and $v \not\equiv 1 \pmod{n}$ there exists a character χ'' modulo n such that $\chi''(v) \neq 1$, then for each w with $(w,N) = 1$ and $w \not\equiv 1 \pmod{N}$ there exists a character χ modulo N such that $\chi(w) \neq 1$.

Proof. Part (i) is a direct consequence of Construction 17.3 and Theorems 17.10, 17.11 and 17.13. For Part (ii), we note that $(w,N) = 1$ implies that $(w,m) = 1$ and $(w,n) = 1$. Also, since $w \not\equiv 1 \pmod{N}$, we have either $w \not\equiv 1 \pmod{m}$ or $w \not\equiv 1 \pmod{n}$. Otherwise, if $w \equiv 1 \pmod{m}$ and $w \equiv 1 \pmod{n}$, then the Chinese Remainder Theorem tells us that $w \equiv 1 \pmod{N}$. Now, if $w \not\equiv 1 \pmod{m}$, we choose $\chi = [\chi', \chi_{n,0}]$ with $\chi'(w) \neq 1$ and use 17.12 to obtain that $\chi(w) = \chi'(w) \neq 1$. Similarly, if $w \not\equiv 1 \pmod{n}$, we choose $\chi = [\chi_{m,0}, \chi'']$ with $\chi''(w) \neq 1$. ■

Theorem 17.15 For any positive integer N, there are exactly $\phi(N)$ characters modulo N. In particular, for any a with $(a,N) = 1$ and $a \not\equiv 1 \pmod{N}$, there always exists a character χ such that $\chi(a) \neq 1$.

Proof. First, there is a unique character modulo 1, namely, $\chi_{1,0}(a) = 1$ for all a. Also, there exists no a such that $a \not\equiv 1 \pmod{1}$. Now we assume that $N \geq 2$. We write N in the canonical form $N = p_1^{\alpha_1}p_2^{\alpha_2}\cdots p_\ell^{\alpha_\ell}$. Using Corollary 17.7 as the base case, we iteratively apply Corollary 17.14 and derive that there are $\phi(p_1^{\alpha_1})\phi(p_2^{\alpha_2})$ characters modulo $p_1^{\alpha_1}p_2^{\alpha_2}$, ..., and $\phi(p_1^{\alpha_1})\phi(p_2^{\alpha_2})\cdots\phi(p_\ell^{\alpha_\ell}) = \phi(N)$ characters modulo N. The second part also holds from this argument. ■

Theorem 17.16 Let N be a positive integer and $\{\chi_0, \ldots, \chi_{\phi(N)-1}\}$ be the set of characters modulo N. Then for any $\chi \in \{\chi_1, \ldots, \chi_{\phi(N)}\}$, the set $\{\chi\chi_1, \ldots, \chi\chi_{\phi(N)-1}\}$ also covers all characters modulo N.

Proof. This follows from the trivial fact that if $\chi\chi_i = \chi\chi_j$, then $\chi_i = \chi_j$. ■

R In general, we may define characters on a group G as group homomorphisms from G to the multiplicative group of a field, usually the field of complex numbers. If G is a finite abelian group, then the number of characters equals the size of G. The Dirichlet characters modulo N correspond to the case of the finite abelian group $(\mathbb{Z}/N\mathbb{Z})^\times$, which is of size $\phi(N)$.

17.4 Orthogonality relations for Dirichlet characters

As we are comfortable with the construction of all Dirichlet characters for a given modulus, we shall establish one of their most important properties, the *orthogonality*.

Theorem 17.17 Let N be a positive integer.

(i) For any Dirichlet character χ modulo N,

$$\sum_{a=1}^{N} \chi(a) = \begin{cases} \phi(N) & \text{if } \chi = \chi_0, \\ 0 & \text{otherwise}, \end{cases} \tag{17.1}$$

where χ_0 is the principal character modulo N;

(ii) For any integer $a \in \mathbb{Z}_{>0}$,

$$\sum_{\chi \bmod N} \chi(a) = \begin{cases} \phi(N) & \text{if } a \equiv 1 \pmod{N}, \\ 0 & \text{otherwise}, \end{cases} \tag{17.2}$$

where the summation runs over all Dirichlet characters modulo N.

Proof. Part (i) is trivial when $\chi = \chi_0$. If $\chi \neq \chi_0$, we choose k with $(k,N) = 1$ and $\chi(k) \neq 1$. By Theorem 3.6, $\{ak : 1 \leq a \leq N\}$ gives a complete system modulo N. Hence,

$$\chi(k) \sum_{a=1}^{N} \chi(a) = \sum_{a=1}^{N} \chi(ak) = \sum_{a=1}^{N} \chi(a),$$

thereby implying the desired result since $\chi(k) \neq 1$.

For Part (ii), it is trivial when $a \equiv 1 \pmod{N}$ or $(a,N) > 1$. If a is such that $(a,N) = 1$ and $a \not\equiv 1 \pmod{N}$, by Theorem 17.15, we have a character $\tilde{\chi}$ modulo N such that $\tilde{\chi}(a) \neq 1$. Also, Theorem 17.16 tells us that if χ run over all characters modulo N, so do $\tilde{\chi}\chi$. Hence,

$$\tilde{\chi}(a) \sum_{\chi \bmod N} \chi(a) = \sum_{\chi \bmod N} \tilde{\chi}\chi(a) = \sum_{\chi \bmod N} \chi(a),$$

which yields the required result since $\tilde{\chi}(a) \neq 1$. ■

As an important consequence of (17.1), we shall show that the partial sum of non-principal characters over positive integers is uniformly bounded.

Corollary 17.18 Let χ be a non-principal character modulo a positive integer N. For $n \geq 1$,

$$\left|\sum_{a=1}^{n} \chi(a)\right| \leq \phi(N). \tag{17.3}$$

Proof. We write $n = qN + r$ where $0 \leq r < N$. Then with recourse to (17.1),

$$\left|\sum_{a=1}^{n} \chi(a)\right| = \left|\left(\sum_{k=0}^{q-1}\sum_{a=1}^{N} \chi(kN+a)\right) + \sum_{a=1}^{r} \chi(qN+a)\right| = \left|\sum_{a=1}^{r} \chi(a)\right| \leq \sum_{a=1}^{r} |\chi(a)| \leq \phi(N),$$

since among $|\chi(1)|, \ldots, |\chi(N-1)|$, there are exactly $\phi(N)$ of them equal to 1, and all others are 0. ■

Theorem 17.19 (Orthogonality Relations). Let N be a positive integer.

(i) For any Dirichlet characters χ_1, χ_2 modulo N,

$$\sum_{a=1}^{N} \chi_1(a)\overline{\chi_2(a)} = \begin{cases} \phi(N) & \text{if } \chi_1 = \chi_2, \\ 0 & \text{otherwise;} \end{cases} \tag{17.4}$$

(ii) For any integers $a_1, a_2 \in \mathbb{Z}_{>0}$,

$$\sum_{\chi \bmod N} \chi(a_1)\overline{\chi(a_2)} = \begin{cases} \phi(N) & \text{if } a_1 \equiv a_2 \pmod{N} \text{ and } (a_1,N) = (a_2,N) = 1, \\ 0 & \text{otherwise,} \end{cases} \tag{17.5}$$

where the summation runs over all Dirichlet characters modulo N.

Proof. For Part (i), we use (17.1) by noting from Theorem 17.2 that $\chi_1(a)\overline{\chi_2(a)} = \chi_1\overline{\chi_2}(a)$ and $\chi_1\overline{\chi_2} = \chi_0$ when $\chi_1 = \chi_2$. For Part (ii), we use (17.2) by noting from Theorem 17.4 that $\chi(a_1)\overline{\chi(a_2)} = \chi(a_1)\chi(\overline{a_2}) = \chi(a_1\overline{a_2})$ whenever a_2 is invertible modulo N. ■

The orthogonality relation (17.5) is by far the most crucial, as it allows one to extract terms indexed by an arithmetic progression from a sum.

Definition 17.4 Let N be a positive integer and a be an integer. We define

$$\mathbf{1}_a(x) = \mathbf{1}_{N,a}(x) := \begin{cases} 1 & \text{if } x \equiv a \pmod{N}, \\ 0 & \text{otherwise.} \end{cases}$$

Corollary 17.20 Let N be a positive integer and a be an integer such that $(a,N) = 1$. Then

$$\mathbf{1}_a(x) = \frac{1}{\phi(N)} \sum_{\chi \bmod N} \overline{\chi(a)}\chi(x), \tag{17.6}$$

where the summation runs over all Dirichlet characters modulo N.

Proof. This is simply (17.5) with $a_1 = x$ and $a_2 = a$. ■

18. Dirichlet's Theorem on primes in arithmetic progressions

18.1 Riemann zeta function

Now we shall take a formal look at the *Riemann zeta function*.

Definition 18.1 For s a complex variable with $\Re(s) > 1$, the *Riemann zeta function* is defined by

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

We have seen from the theory of Dirichlet series that the series $\sum_{n\geq 1} \frac{1}{n^s}$ is absolutely convergent in the half-plane $\sigma > 1$, and the Riemann zeta function is analytic in its domain. Further, we have the Euler product for $\zeta(s)$ by taking $f = \mathbf{1}$ in (16.19).

Theorem 18.1 For $\sigma > 1$,

$$\zeta(s) = \prod_p \frac{1}{1-p^{-s}}. \tag{18.1}$$

As for many other functions of a complex variable, we also hope to extend the domain of definition of the Riemann zeta function through analytic continuation.

Theorem 18.2 For $s \neq 1$ with $\sigma > 0$,

$$\zeta(s) = \frac{s}{s-1} - s\int_1^\infty \frac{\{u\}}{u^{s+1}}du. \tag{18.2}$$

In particular, $\zeta(s)$ is analytic in the half-plane $\sigma > 0$ but with a simple pole at $s = 1$, with residue 1.

Proof. We start with the case $\sigma > 1$. Note that by Euler's summation formula (15.2),

$$\begin{aligned}\sum_{n\leq x} \frac{1}{n^s} &= \int_{1^-}^x \frac{du}{u^s} - s\int_{1^-}^x \frac{\{u\}}{u^{s+1}}du + 1 - \frac{\{x\}}{x^s}\\ &= \frac{s}{s-1} - \frac{x^{1-s}}{s-1} - \frac{\{x\}}{x^s} - s\int_1^x \frac{\{u\}}{u^{s+1}}du.\end{aligned}$$

Now for $\sigma > 1$, we have $\frac{x^{1-s}}{s-1} \to 0$ and $\frac{\{x\}}{x^s} \to 0$ as $x \to \infty$. Also, the integral $\int_1^\infty \{u\}u^{-s-1}du$ is convergent in the half-plane $\sigma > 0$ by comparison with $\int_1^\infty u^{-\sigma-1}du$. Hence, (18.2) is valid for $\sigma > 1$. Further, the convergence of the aforementioned integral is also locally uniform in the open half-plane $\sigma > 0$, and by Weierstrass's theorem on uniformly convergent sequences of analytic functions, we see that $\int_1^\infty \{u\}u^{-s-1}du$ is an analytic function in this half-plane. Recalling the uniqueness of analytic continuation, the formula (18.2) is valid for $\sigma > 0$ with $s \neq 1$, and at $s = 1$, we have a simple pole from $\frac{s}{s-1}$. ■

R In general, the Riemann zeta function can be analytically continued to $\mathbb{C}\backslash\{1\}$. This can be achieved by applying the Euler–Maclaurin formula, which extends Euler's summation formula (15.2) via repeated integration by parts. A more immediate way is by invoking the functional equation for the Riemann zeta function

$$\zeta(1-s) = 2(2\pi)^{-s}\Gamma(s)\cos(\tfrac{\pi s}{2})\zeta(s),$$

where $\Gamma(s)$ is the Gamma function; see T. M. Apostol, Ch. 12.

18.2 Dirichlet L-functions

Noting that the Dirichlet characters are arithmetic functions, we may further consider their associated Dirichlet series.

Definition 18.2 Let χ be a Dirichlet character modulo N. Its Dirichlet series

$$L(s,\chi) := \sum_{n\geq 1} \frac{\chi(n)}{n^s}$$

is called the *Dirichlet L-function*, or the *Dirichlet L-series*, associated with χ.

Recall that we have $\chi(n) = 0$ if $(n,N) > 1$ by definition, and $|\chi(n)| = 1$ if $(n,N) = 1$ by Theorem 17.1. Therefore, the series $\sum_{n\geq 1}\frac{\chi(n)}{n^s}$ is absolutely convergent in the half-plane $\sigma > 1$. Noting that χ is completely multiplicative, we then derive the Euler product for $L(s,\chi)$ by taking $f = \chi(n)$ in (16.19).

Theorem 18.3 Let χ be a Dirichlet character modulo N. For $\sigma > 1$,

$$L(s,\chi) = \prod_p \frac{1}{1-\chi(p)p^{-s}}. \tag{18.3}$$

Now we work on the analytic properties of Dirichlet L-functions.

Theorem 18.4 Let χ_0 be the principal Dirichlet character modulo N. For $s \neq 1$ with $\sigma > 0$,

$$L(s,\chi_0) = \zeta(s)\prod_{p|N}\left(1-\frac{1}{p^s}\right). \tag{18.4}$$

In particular, $L(s,\chi_0)$ is analytic in the half-plane $\sigma > 0$ but with a simple pole at $s = 1$, with residue $\phi(N)/N$.

Proof. It is sufficient to prove (18.4) for $\sigma > 1$; we may then analytically extend the domain to the larger half-plane $\sigma > 0$ as $\prod_{p|N}\left(1-\frac{1}{p^s}\right)$ is entire. Since χ_0 is principal, we have $\chi_0(p) = 1$ if $p \nmid N$, and 0 if $p \mid N$. For $\sigma > 1$, we derive from the Euler products for $\zeta(s)$

and $L(s,\chi_0)$ that

$$L(s,\chi_0)=\prod_{p\nmid N}\frac{1}{1-p^{-s}}=\prod_{p}\frac{1}{1-p^{-s}}\prod_{p|N}\left(1-p^{-s}\right)=\zeta(s)\prod_{p|N}\left(1-p^{-s}\right),$$

as required. The simple pole of $L(s,\chi_0)$ at $s=1$ comes from the simple pole of $\zeta(s)$. Also, the residue at $s=1$ equals $\prod_{p|N}\left(1-p^{-1}\right)=\frac{\phi(N)}{N}$. ■

Theorem 18.5 Let $\chi\neq\chi_0$ be a non-principal Dirichlet character modulo N. The series $\sum_{n\geq 1}\chi(n)n^{-s}$ converges in the half-plane $\sigma>0$. Also, $L(s,\chi)$ is analytic in this half-plane, and

$$L(s,\chi)=s\int_1^{\infty}\frac{X(u)}{u^{s+1}}du, \tag{18.5}$$

where $X(u):=\sum_{n\leq u}\chi(n)$.

Proof. We apply Abel's summation formula (15.1) and derive that

$$\sum_{n\leq x}\frac{\chi(n)}{n^s}=\frac{X(x)}{x^s}+s\int_1^{x}\frac{X(u)}{u^{s+1}}du.$$

Recall from Corollary 17.18 that $|X(u)|\leq\phi(N)$ for all $u>0$. Hence, for $\sigma>0$, $X(x)x^{-s}\to 0$ as $x\to\infty$. Also, the integral $\int_1^{\infty}X(u)u^{-s-1}du$ is convergent in the half-plane $\sigma>0$ by comparison with $\phi(N)\int_1^{\infty}u^{-\sigma-1}du$. The above argument implies the convergence of $\sum_{n\geq 1}\chi(n)n^{-s}$ for $\sigma>0$, and also the formula (18.5). Finally, $L(s,\chi)$ is analytic in the half-plane $\sigma>0$ due to the Analyticity Theorem in Rule 16.5. ■

One crucial property of the Dirichlet L-function is the non-vanishing of $L(1,\chi)$ for all non-principal characters χ.

Theorem 18.6 Let $\chi\neq\chi_0$ be a non-principal Dirichlet character modulo N. Then $L(1,\chi)\neq 0$.

We shall separate this theorem into two parts, according to whether χ is non-real or real; see Theorem 18.8 and 18.9, respectively.

Here one necessary step is to consider the logarithm of the L-functions. Since the L-functions are complex-valued, we should choose a suitable branch for the complex logarithm. Throughout, what we mean by log is the principal branch, which is analytically continued by the normal real logarithm to $\mathbb{C}\backslash\mathbb{R}_{\leq 0}$. For this choice of log, we know that $\log z$ is real if and only if z is a positive real number. Also, it has the power series expansion for $|z|<1$,

$$\log\frac{1}{1-z}=\sum_{m\geq 1}\frac{z^m}{m}.$$

Now, let us constrain our focus from the whole half-plane $\sigma>0$ to the positive real axis.

Lemma 18.7 Let N be a positive integer, and define

$$Z(s):=\prod_{\chi \bmod N}L(s,\chi).$$

Then for $\sigma > 1$, $Z(\sigma)$ is real-valued with $Z(\sigma) > 1$.

Proof. We make use of the Euler product (18.3) for L-functions, and obtain that

$$\begin{aligned}\log Z(\sigma) = \log \prod_{\chi \bmod N} \prod_{p} \frac{1}{1-\chi(p)p^{-\sigma}} &= \sum_{\chi \bmod N} \sum_{p} \log \frac{1}{1-\chi(p)p^{-\sigma}} \\ = \sum_{\chi \bmod N} \sum_{p} \sum_{m\geq 1} \frac{\chi(p)^m}{mp^{m\sigma}} &= \sum_{p} \sum_{m\geq 1} \frac{1}{mp^{m\sigma}} \sum_{\chi \bmod N} \chi(p)^m \\ = \sum_{p} \sum_{m\geq 1} \frac{1}{mp^{m\sigma}} \sum_{\chi \bmod N} \chi(p^m).\end{aligned}$$

By (17.2),

$$\sum_{\chi \bmod N} \chi(p^m) = \begin{cases} \phi(N) & \text{if } p^m \equiv 1 \pmod{N}, \\ 0 & \text{otherwise.} \end{cases}$$

For every $p \nmid N$, we may always find such m with $p^m \equiv 1 \pmod{N}$, viz. m are multiples of $\operatorname{ord}_N p$. It follows that $\log Z(\sigma)$ is a positive real number, and hence $Z(\sigma)$ is real-valued with $Z(\sigma) > 1$. ■

Theorem 18.8 Let $\chi \neq \chi_0$ be a non-principal non-real Dirichlet character modulo N. Then $L(1,\chi) \neq 0$.

Proof. We argue by contradiction. If $\chi_\dagger$ is a non-principal non-real character such that $L(1,\chi_\dagger) = 0$, so is $\overline{\chi_\dagger}$ since

$$L(1,\overline{\chi_\dagger}) = \sum_{n\geq 1} \frac{\overline{\chi_\dagger(n)}}{n} = \overline{\sum_{n\geq 1} \frac{\chi_\dagger(n)}{n}} = \overline{L(1,\chi_\dagger)} = 0.$$

Note that the three characters χ_0, $\chi_\dagger$ and $\overline{\chi_\dagger}$ are pairwise distinct. Hence, we may rewrite $Z(\sigma)$ in Lemma 18.7 as

$$Z(\sigma) = L(\sigma,\chi_0)L(\sigma,\chi_\dagger)L(\sigma,\overline{\chi_\dagger}) \prod_{\chi \neq \chi_0,\chi_\dagger,\overline{\chi_\dagger}} L(\sigma,\chi).$$

If we look at a small neighborhood $|\sigma - 1| < \varepsilon \to 0$ near $\sigma = 1$, we have that $L(1+\varepsilon,\chi_0) = O(\varepsilon^{-1})$ by the simple pole of $L(s,\chi_0)$ at $s = 1$, and that $L(1+\varepsilon,\chi_\dagger) = O(\varepsilon)$, $L(1+\varepsilon,\overline{\chi_\dagger}) = O(\varepsilon)$ and $L(1+\varepsilon,\chi) = O(1)$ otherwise by the analyticity of L-functions in the half-plane $\sigma > 0$ for every non-principal characters together with the assumption that $L(1,\chi_\dagger) = L(1,\overline{\chi_\dagger}) = 0$. Hence, as $\varepsilon \to 0$, $Z(1+\varepsilon) = O(\varepsilon) \to 0$. But this violates Lemma 18.7, claiming that $Z(1+\varepsilon) > 1$ whenever $\varepsilon > 0$. ■

However, for non-principal real characters, we *cannot* proceed with the above argument as the complex conjugate of a real character is still itself. So we cannot automatically get another copy of $L(1+\varepsilon,\chi_\dagger') = O(\varepsilon)$ as above to reduce $Z(1+\varepsilon)$ to $O(\varepsilon)$. Now we shall adopt an elegant device due to Paul Monsky (*Amer. Math. Monthly* **100** (1993), no. 9, 861–862).

Theorem 18.9 Let $\chi \neq \chi_0$ be a non-principal real Dirichlet character modulo N. Then $L(1,\chi) \neq 0$.

Proof. We start with the Lambert series with $0<x<1$,

$$F(x)=\sum_{k\geq 1}\frac{\chi(k)x^k}{1-x^k}.$$

Note that this series is absolutely convergent by comparison with $\sum_{k\geq 1}\frac{x^k}{1-x^k}$, whose convergence follows by the ratio test. We may also expand $F(x)$ as a power series in x, say, $F(x)=\sum_{n\geq 1}a(n)x^n$. With recourse to (12.3), we have

$$a(n)=\sum_{d|n}\chi(d).$$

By Theorem 14.9, the arithmetic function a is multiplicative since $a=\chi*\mathbf{1}$ while both χ and $\mathbf{1}$ are multiplicative. We then claim that $a(n)\geq 0$ for all n. Here it is sufficient to verify that $a(p^\alpha)\geq 0$ for all prime powers p^α. Recalling that χ is real, and hence that $\chi(p)\in\{-1,0,1\}$, we have

$$a(p^\alpha)=1+\chi(p)+\cdots+\chi(p^\alpha)=1+\chi(p)+\cdots+\chi(p)^\alpha\geq 0,$$

by grouping terms $\chi(p)^{2j}+\chi(p)^{2j+1}\geq 0$. In particular, we deduce by the same argument that $a(p^{2\beta})\geq 1$ for all even powers of primes $p^{2\beta}$. Since a is multiplicative, we have $a(n^2)\geq 1$ for all n. It follows from the above that the series $\sum_{n\geq 1}a(n)$ diverges. For each $M\geq 1$, we have $\limsup_{x\to 1^-}F(x)\geq \lim_{x\to 1^-}\sum_{n\leq M}a(n)x^n=\sum_{n\leq M}a(n)$. Hence, as $x\to 1^-$, $F(x)\to\infty$.

Let us assume that the real character $\chi\neq\chi_0$ is such that $L(1,\chi)=0$. Then

$$-F(x)=\frac{L(1,\chi)}{1-x}-F(x)=\sum_{n\geq 1}\chi(n)\left(\frac{1}{(1-x)n}-\frac{x^n}{1-x^n}\right)=:\sum_{n\geq 1}f_n(x)\chi(n).$$

Then for $0<x<1$, $\lim_{n\to\infty}f_n(x)=0$. Also,

$$\begin{aligned}f_n(x)-f_{n+1}(x)&=\left(\frac{1}{(1-x)n}-\frac{x^n}{1-x^n}\right)-\left(\frac{1}{(1-x)(n+1)}-\frac{x^{n+1}}{1-x^{n+1}}\right)\\&=\frac{1}{1-x}\left(\frac{1}{n(n+1)}-\frac{x^n(1-x)^2}{(1-x^n)(1-x^{n+1})}\right).\end{aligned}$$

Further, we apply the arithmetic-geometric mean inequality and find that for every positive integer k,

$$\frac{1-x^k}{1-x}=1+x+\cdots+x^{k-1}=\frac{1+x^{k-1}}{2}+\frac{x+x^{k-2}}{2}+\cdots+\frac{x^{k-1}+1}{2}\geq kx^{\frac{k-1}{2}}.$$

Hence,

$$f_n(x)-f_{n+1}(x)\geq\frac{1}{1-x}\left(\frac{1}{n(n+1)}-\frac{x^{\frac{1}{2}}}{n(n+1)}\right)>0.$$

That is, $f_n(x)$ is a decreasing sequence whenever $0<x<1$.

Now, if we as usual write $X(u)=\sum_{n\leq u}\chi(n)$, then for every integer $M\geq 1$, by replacing $\chi(n)$ with $X(n)-X(n-1)$ and then rearranging terms, we have

$$\sum_{n\leq M}f_n(x)\chi(n)=X(M)f_{M+1}(x)+\sum_{n\leq M}X(n)\big(f_n(x)-f_{n+1}(x)\big).$$

Since $f_n(x) \searrow 0$ as $n \to \infty$, we have the following bound by also recalling from Corollary 17.18 that $|X(u)| \leq \phi(N)$,

$$\left|\sum_{n\leq M} f_n(x)\chi(n)\right| \leq \phi(N) f_{M+1}(x) + \phi(N) \sum_{n\leq M} \big(f_n(x) - f_{n+1}(x)\big) = \phi(N) f_1(x),$$

for all $0 < x < 1$. However, $f_1(x) = \frac{1}{1-x} - \frac{x}{1-x} = 1$. Hence, we have that $|F(x)| \leq \phi(N)$ whenever $0 < x < 1$. But this contradicts what we have shown earlier that $F(x) \to \infty$ as $x \to 1^-$. Thus, we cannot have any real character $\chi \neq \chi_0$ with $L(1,\chi) = 0$. ■

R The above evaluation of $\sum_{n\leq M} f_n(x)\chi(n)$ is indeed an instance of the *Abel transformation*, also known as *summation by parts*.

Theorem 18.10 (Abel Transformation). For sequences $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, if we put $A_n := \sum_{N<m\leq n} a_m$, then for integers $0 \leq N < M$,

$$\sum_{N<n\leq M} a_n b_n = A_M b_{M+1} + \sum_{N<n\leq M} A_n (b_n - b_{n+1}).$$

This can be regarded as a *discrete* version of Abel's summation formula (15.1). For its proof, we simply replace a_n with $A_n - A_{n-1}$ and rearrange terms.

18.3 Dirichlet's Theorem on primes in arithmetic progressions

Now we are in a position to put the finishing touches to the proof of Dirichlet's theorem on primes in arithmetic progressions.

Theorem 18.11 (Dirichlet's Theorem on Primes in Arithmetic Progressions). There are infinitely many primes congruent to a modulo N provided that $(a,N) = 1$.

We first summarize what was obtained earlier.

Lemma 18.12 In the half-plane $\sigma > 1$, we have, as $s \to 1$,

$$\log L(s,\chi_0) \sim -\log(s-1), \tag{18.6}$$

where χ_0 is the principal Dirichlet character modulo a positive integer N, and

$$\log L(s,\chi) = O(1), \tag{18.7}$$

where $\chi \neq \chi_0$ is a non-principal Dirichlet character modulo N.

Proof. For (18.6), we use (18.2) and (18.4). For (18.7), we know from Theorems 18.5 and 18.6 that in the half-plane $\sigma > 0$ there exists a neighborhood near $s = 1$ such that within this neighborhood $L(s,\chi)$ is bounded and $L(s,\chi) \neq 0$, thereby implying that $\log L(s,\chi)$ is also bounded. ■

Proof of Theorem 18.11. We make use of the Euler product (18.3) for L-functions, and obtain that for $\sigma > 1$,

$$\begin{aligned}\sum_{\chi \bmod N} \overline{\chi(a)} \log L(s,\chi) &= \sum_{\chi \bmod N} \overline{\chi(a)} \sum_p \log \frac{1}{1-\chi(p)p^{-s}}\\ &= \sum_{\chi \bmod N} \overline{\chi(a)} \sum_p \sum_{m\geq 1} \frac{\chi(p)^m}{mp^{ms}}\end{aligned}$$

$$=\sum_p \sum_{\chi \bmod N} \overline{\chi(a)}\frac{\chi(p)}{p^s} + \sum_p \sum_{m\geq 2} \sum_{\chi \bmod N} \overline{\chi(a)}\frac{\chi(p)^m}{mp^{ms}}.$$

For the first term, we deduce from Corollary 17.20 that

$$\sum_p \sum_{\chi \bmod N} \overline{\chi(a)}\frac{\chi(p)}{p^s} = \phi(N) \sum_{p\equiv a \bmod N} \frac{1}{p^s}.$$

For the second term, we have the bound for $\sigma > 1$,

$$\begin{aligned}\left|\sum_p \sum_{m\geq 2} \sum_{\chi \bmod N} \overline{\chi(a)}\frac{\chi(p)^m}{mp^{ms}}\right| &\leq \sum_p \sum_{m\geq 2} \sum_{\chi \bmod N} \frac{1}{mp^{m\sigma}} \\ &< \frac{\phi(N)}{2}\sum_p \sum_{m\geq 2} \frac{1}{p^m} = \frac{\phi(N)}{2}\sum_p \frac{1}{p(p-1)} \\ &< \frac{\phi(N)}{2}\sum_{n\geq 2} \frac{1}{n(n-1)} = \frac{\phi(N)}{2}.\end{aligned}$$

Finally, since $(a,N)=1$, we have $\overline{\chi_0(a)}=1$. It follows from (18.6) and (18.7) that, as $s \to 1$ in the half-plane $\sigma > 1$,

$$\sum_{\chi \bmod N} \overline{\chi(a)} \log L(s,\chi) \sim -\log(s-1) \to \infty.$$

This in turn implies the divergence of $\sum_{p\equiv a \bmod N} \frac{1}{p}$, and therefore the infinitude of primes congruent to a modulo N. ■

19. Rational and irrational numbers

19.1 Algebraic structures

We begin with some basic concepts in abstract algebra: *groups*, *rings* and *fields*.

Definition 19.1 A *group* $(G,\circ)$ is a finite or infinite set G of elements together with a binary operation "$\circ$", called the *group operation*, such that the following properties hold:

(i) *Closure*: For all $a,b \in G$, we have $a \circ b \in G$;

(ii) *Associativity*: For all $a,b,c \in G$, we have $(a \circ b) \circ c = a \circ (b \circ c)$;

(iii) *Identity*: There exists an *identity element* $e \in G$ such that for all $a \in G$, we have $a \circ e = e \circ a = a$;

(iv) *Inverse*: For all $a \in G$, there exists an element $b \in G$, which is called the *inverse* of a, such that $a \circ b = b \circ a = e$; the inverse of a is usually denoted by a^{-1} or $-a$.

Strictly speaking, a group and its underlying set are two different mathematical objects as in a group we must have an associated binary operation. However, when this group operation is clear, we often abuse notation and use the name of the set G to represent the group.

Example 19.1 Considering the set $\mathbb{R}$ of real numbers, we have a group $(\mathbb{R},+)$ under the usual addition "$+$", where 0 is the identity element and the opposite of a, namely, $-a$, is the inverse of $a \in \mathbb{R}$. But in most cases, we simply write $\mathbb{R}$ for this group. Also, considering the set $\mathbb{R}^\times$ of nonzero real numbers, we have a group $(\mathbb{R}^\times,\times)$ under the usual multiplication "$\times$", where 1 is the identity element and the reciprocal of a, namely, $1/a$ or a^{-1}, is the inverse of $a \in \mathbb{R}^\times$. We may write $\mathbb{R}^\times$ for this group as well. ■

R In general, we speak of an *additive group* whenever the group operation is notated as addition, and in this case, the identity element is typically denoted by 0 and the inverse of an element a is denoted by $-a$. Similarly, we speak of a *multiplicative group* whenever the group operation is notated as multiplication, and in this case, the identity element is typically denoted by 1 and the inverse of an element a is denoted by a^{-1}.

Definition 19.2 A group $(G,\circ)$ is *abelian* if the following property further holds:

(v) *Commutativity*: For all $a,b \in G$, we have $a \circ b = b \circ a$.

R It is a common convention that for an abelian group we may use either additive or multiplicative notation, but for a nonabelian group only multiplicative notation is used.

If the requirement of having an inverse for every element is removed from Definition 19.1, we arrive at a *monoid*.

Definition 19.3 A *monoid* is a set which is **closed** under an **associative** binary operation and has an **identity element**.

Example 19.2 The set of integers $\mathbb{Z}$ under the usual multiplication forms a monoid, and 1 is the identity element. However, $\mathbb{Z}$ under the usual addition forms a group with 0 the identity element. ■

Definition 19.4 A *ring* $(R,+,\cdot)$ is a finite or infinite set R of elements together with two binary operations "$+$" (addition) and "$\cdot$" (multiplication) such that the following properties hold:

(i) R is an abelian group under addition with *addition identity* 0;
(ii) R is a monoid under multiplication with *multiplication identity* 1;
(iii) Multiplication is *distributive* with respect to addition, i.e. for all $a,b,c \in R$, we have $a\cdot(b+c) = a\cdot b + a\cdot c$ and $(b+c)\cdot a = b\cdot a + c\cdot a$.

Example 19.3 The set of integers $\mathbb{Z}$ under the usual addition and multiplication forms a ring. ■

Definition 19.5 If the ring multiplication of $(R,+,\cdot)$ is commutative, that is, $a\cdot b = b\cdot a$ for all $a,b \in R$, then R is called a *commutative ring*.

Example 19.4 The ring of integers, $\mathbb{Z}$, is commutative. However, the ring of 2×2 matrices over integers, $\mathrm{Mat}_{2,2}(\mathbb{Z})$, under the usual matrix addition and multiplication is noncommutative. ■

Definition 19.6 A nonzero commutative ring in which the product of any two nonzero elements is nonzero is called an *integral domain*.

Example 19.5 The quotient ring $\mathbb{Z}/p\mathbb{Z}$ with p a prime is an integral domain. However, $\mathbb{Z}/m\mathbb{Z}$ with m a composite is not an integral domain. For instance, 2 and 3 in $\mathbb{Z}/6\mathbb{Z}$ are nonzero, but $2\times 3 = 6 \equiv 0 \pmod 6$, and hence $2\times 3 = 0$ in $\mathbb{Z}/6\mathbb{Z}$. ■

Proposition 19.1 Let R be an integral domain. If $a,b,c \in R$ are such that $a \neq 0$ and $ab = ac$, then $b = c$.

Proof. We know from $ab = ac$ that $a(b-c) = 0$. If $b \neq c$, or $b - c \neq 0$, then by $a \neq 0$, we have $a(b-c) \neq 0$ since R is an integral domain. This leads to a contradiction. ■

Definition 19.7 A *field* $(F,+,\cdot)$ is a commutative ring where $0 \neq 1$ and all nonzero elements are invertible under multiplication. That is, the following properties hold for all $a,b,c \in F$:

(i) *Closure of addition and multiplication*: $a+b \in F$ and $a\cdot b \in F$;
(ii) *Associativity of addition and multiplication*: $(a+b)+c = a+(b+c)$ and $(a\cdot b)\cdot c = a\cdot(b\cdot c)$;
(iii) *Commutativity of addition and multiplication*: $a+b = b+a$ and $a\cdot b = b\cdot a$;
(iv) *Additive and multiplicative identity*: There exist two **different** elements 0 and 1

in F such that $a+0=a$ and $a\cdot 1=a$;

(v) *Additive inverses*: For every $a\in F$, there exists an element in F, denoted by $-a$, such that $a+(-a)=0$; here $-a$ is called the *additive inverse* of a;

(vi) *Multiplicative inverses*: For every $a\in F$ with $a\neq 0$, there exists an element in F, denoted by a^{-1}, such that $a\cdot a^{-1}=1$; here a^{-1} is called the *multiplicative inverse* of a;

(vii) *Distributivity of multiplication over addition*: $a\cdot(b+c)=a\cdot b+a\cdot c$.

Example 19.6 The set of real numbers $\mathbb{R}$ under the usual addition and multiplication forms a field. Also, the set of complex numbers $\mathbb{C}$ under the usual addition and multiplication forms a field. ■

19.2 Rational and irrational numbers

Definition 19.8 A *rational number* or a *rational* x is a real number that can be expressed as the quotient $x=\frac{a}{b}$ of integers a and b with $b\neq 0$. In particular, if $(a,b)=1$ and $b>0$, then $\frac{a}{b}$ is called the *irreducible expression* of x.

Theorem 19.2 Every rational number has a unique irreducible expression.

Proof. The existence of an irreducible expression for any rational x comes from the fact that if $x=\frac{a}{b}$ with $d=(a,b)$ and $b>0$, then

$$x=\frac{a'}{b'}$$

where $a'=a/d$ and $b'=b/d$ so that $(a',b')=1$. For uniqueness, if x has irreducible expressions

$$x=\frac{a_1}{b_1}=\frac{a_2}{b_2}$$

with $(a_1,b_1)=(a_2,b_2)=1$ and $b_1,b_2>0$, then $a_1b_2=a_2b_1$. Now $b_1\mid a_1b_2$ and $(a_1,b_1)=1$ imply that $b_1\mid b_2$ by Theorem 2.6. Similarly, we have $b_2\mid b_1$, and thus conclude that $b_1=b_2$, which further yields $a_1=a_2$. ■

Theorem 19.3 The set of rational numbers $\mathbb{Q}$ under the usual addition and multiplication forms a field.

Proof. This statement follows readily by verifying the properties in Definition 19.7. ■

Definition 19.9 An *irrational number* or an *irrational* is a real number that is not rational, i.e. it cannot be expressed as a quotient $\frac{a}{b}$ of integers a and b with $b\neq 0$.

R The set of irrational numbers under either the usual addition or multiplication is **not** a group. This is because neither 0 nor 1 is irrational.

It is in general hard to determine if a real number is irrational or not. Among the known results about irrationality, we know that every non-integral radical of a positive integer is irrational, and that every rational power of e is irrational except $e^0=1$; the two results will be proved in the next two sections, respectively.

For other results, we have the irrationality of all rational powers of π, again except $\pi^0=1$; this fact is a direct implication of the transcendence of π, but we will not cover it in the current series of notes and the interested reader may refer to Hardy and Wright's

Introduction, Sect. 11.14. Now by the fact (see Apostol, Sect. 12.12) that for $n \geq 1$, the Riemann zeta function has the following evaluation (with B_{2n} the $2n$-th Bernoulli number),

$$\zeta(2n) = \frac{(-1)^{n+1} B_{2n} (2\pi)^{2n}}{2(2n)!}, \tag{19.1}$$

which is a rational multiple of π^{2n}, we know that $\zeta(2n)$ is irrational for every positive integer n. Another important result is the irrationality of $\zeta(3)$, which was proved by the French mathematician Roger Apéry (*Astérisque* **61** (1979), 11–13).

However, it remains an open problem for the irrationality of $\zeta(2n+1)$ with $n \geq 2$ an integer. Also, it is unknown if $\pi + e$, πe, π/e, π^e, 2^e, $\log \pi$ or the Euler–Mascheroni constant γ are irrational.

19.3 Irrationality of radicals

The very first study of irrationality dates back to ancient Greece. Here, we start with a result that is usually attributed to a Pythagorean, possibly Hippasus of Metapontum.

Theorem 19.4 $\sqrt{2}$ is irrational.

Hippasus's reasoning is as follows. First, we assume that $\sqrt{2}$ is rational, and hence by Theorem 19.2, we can write $\sqrt{2}$ in the irreducible expression

$$\sqrt{2} = \frac{a}{b},$$

where $(a,b) = 1$ and $b > 0$. Then

$$a^2 = 2b^2,$$

thereby implying that a is even, and thus that a^2 is a multiple of 4. Then b^2 is also even, yielding that b is even. But this violates the assumption that $(a,b) = 1$.

The above argument is of course intuitive. However, to ensure the possibility of further generalizations, we shall go beyond the consideration for parity.

Proof. First, we know that $1 < \sqrt{2} < 2$, and hence that $\sqrt{2}$ is not an integer. Thus, if we assume that $\sqrt{2}$ is rational and write $\sqrt{2}$ in the irreducible expression $\sqrt{2} = \frac{a}{b}$, where $(a,b) = 1$ and $b > 0$ as before, then $b > 1$. In other words, b has a prime factor p. Now, $p \mid 2b^2 = a^2$, and by Corollary 2.7, $p \mid a$. This violates the assumption that $(a,b) = 1$. ■

Now we go ahead with a more general result.

Theorem 19.5 Let x be a real number such that

$$x^m + c_{m-1} x^{m-1} + \cdots + c_1 x + c_0 = 0 \tag{19.2}$$

with all coefficients c_i integers. Then either x is an integer or x is irrational.

Proof. If x is an integer, then we are done. Now assume that x is a non-integral rational; so written in the irreducible expression $x = \frac{a}{b}$ with $(a,b) = 1$, we have $b > 1$. Note that (19.2) is equivalent to

$$a^m + c_{m-1} a^{m-1} b + \cdots + c_1 a b^{m-1} + c_0 b^m = 0.$$

Thus, if p is a prime factor of b, then $p \mid a^m$, and hence $p \mid a$ by repeatedly using Corollary 2.7. This leads to a contradiction. ■

■ **Example 19.7** *For any positive integers m and N, the number $\sqrt[m]{N}$ is irrational unless N is the m-th power of a certain positive integer n. In particular, if $N > 1$ is such that there is a prime p with $p \mid N$ and $p^m \nmid N$, then $\sqrt[m]{N}$ is irrational.* Here we note that $x = \sqrt[m]{N}$ is such that $x^m - N = 0$. ■

■ **Example 19.8** *The number $\sqrt{2} + \sqrt{3}$ is irrational.* Here we note that $x = \sqrt{2} + \sqrt{3}$ is such that $x^4 - 10x^2 + 1 = 0$. ■

19.4 Irrationality of e

We start with the irrationality of e, whose proof is nearly elementary.

Theorem 19.6 e is irrational.

Proof. We shall make use of the following infinite series representation of e, which may be deduced from the Taylor expansion of e^x:

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \cdots.$$

Assume that e is rational, and write $e = \frac{a}{b}$ with a and b integers and $b > 0$. Further, let n be a positive integer with $n \ge b$. Then $b \mid n!$ and hence the number

$$S := n!\left(e - 1 - \frac{1}{1!} - \frac{1}{2!} - \cdots - \frac{1}{n!}\right)$$

is an integer. However, we also have

$$\begin{aligned} 0 < S &= n!\left(\frac{1}{(n+1)!} + \frac{1}{(n+2)!} + \frac{1}{(n+3)!} + \cdots\right) \\ &< \frac{1}{n+1} + \frac{1}{(n+1)^2} + \frac{1}{(n+1)^3} + \cdots \\ &= \frac{1}{n} \le 1. \end{aligned}$$

Hence, S cannot be an integer and we arrive at a contradiction. ■

For the irrationality of rational powers of e other than $e^0 = 1$, we require a clever idea due to the French mathematician Charles Hermite (*Comptes rendus de l'Académie des Sciences (Paris)* **77** (1873), 18–24) that requires some calculus. However, the underlying logic is similar — We assume that the statement is false, and then construct a quantity that is an integer but also falls into the open interval $(0,1)$, which is definitely absurd.

Lemma 19.7 Let n be a positive integer and define

$$f(x) = \frac{x^n(1-x)^n}{n!}. \tag{19.3}$$

Then

(i) The function $f(x)$ is a polynomial in x of the form $f(x) = \frac{1}{n!}\sum_{i=n}^{2n} c_i x^i$, where the coefficients c_i are integers;

(ii) For $0 < x < 1$, we have $0 < f(x) < \frac{1}{n!}$;

(iii) The derivatives $f^{(k)}(0)$ and $f^{(k)}(1)$ are integers for all $k \ge 0$.

Proof. The first two parts are plain. For Part (iii), it is immediate that $f^{(k)}(0) = 0$ when $k < n$ or $k > 2n$. For the cases where $n \le k \le 2n$, we have $f^{(k)}(0) = \frac{k!}{n!}c_k$, which is an integer by first noting that $\frac{k!}{n!}$ is an integer since $k \geq n$ and then recalling from Part (i) that c_k is also an integer. Finally, we have $f^{(k)}(1) = (-1)^k f^{(k)}(0)$ since $f(x) = f(1-x)$. ■

Theorem 19.8 e^u is irrational for every rational u with $u \neq 0$.

Proof. We first prove that e^m is irrational for every positive integer m. Assume not, and write $e^m = \frac{a}{b}$ with a and b positive integers. Recalling (19.3), we define

$$F(x) := m^{2n} f(x) - m^{2n-1} f'(x) + m^{2n-2} f''(x) - \cdots + f^{(2n)}(x),$$

where n is chosen so that

$$n! > am^{2n}.$$

Now the fact we shall use is that

$$\frac{d}{dx}\left[e^{mx} F(x)\right] = m^{2n+1} e^{mx} f(x),$$

where we note that $f^{(2n+1)}(x)$ vanishes for all x. Hence,

$$S := b \int_0^1 m^{2n+1} e^{mx} f(x) dx = b\left[e^{mx} F(x)\right]_0^1 = aF(1) - bF(0)$$

is an integer by Lemma 19.7(iii). However, we also have, with recourse to Lemma 19.7(ii), that

$$0 < S = b \int_0^1 m^{2n+1} e^{mx} f(x) dx < \frac{bm^{2n} e^m}{n!} = \frac{am^{2n}}{n!} < 1,$$

where the last inequality comes from our choice of n. Therefore, we are led to a contradiction, thereby implying that e^m is irrational for every positive integer m. Finally, if there exists a certain rational $u = \frac{s}{t} \neq 0$ (without loss of generality, $s > 0$ is assumed) such that e^u is rational, then so is $(e^u)^t = e^s$ since $\mathbb{Q}$ is a field. However, this contradicts what we have proved earlier that e^s must be irrational. ■

20. Algebraic and transcendental numbers

20.1 Fundamental theorem of algebra

We have seen in the previous lecture that every rational number is a root of a linear polynomial with integer coefficients. Now our attention is turned to the roots of generic polynomials, and we shall witness the *Fundamental Theorem of Algebra*, one of the milestones in the history of mathematics.

Theorem 20.1 (Fundamental Theorem of Algebra). Every nonzero single-variable polynomial of degree $n \geq 1$ with complex coefficients has, counted with multiplicity, exactly n complex roots.

Our proof of this theorem only relies on three facts from basic calculus:

▷ Polynomial functions are continuous.

▷ For any complex number z with $|z| = 1$ and any positive integer m, there exists a complex number ζ with $|\zeta| = 1$ such that $\zeta^m = z$. In fact, if z is written as $z = e^{2\pi i\theta}$ for a certain real θ, then ζ can be chosen as $\zeta = e^{\frac{2\pi i\theta}{m}}$.

▷ *Cauchy's minimum principle*: A continuous real-valued function f on a compact set S assumes a minimum in S.

The first two assertions lead us to a key result, which is now known as the *d'Alembert–Argand Lemma*, due to the French mathematician Jean le Rond d'Alembert and the Swiss amateur mathematician Jean-Robert Argand.

Lemma 20.2 (d'Alembert–Argand Lemma). Let $p(z) = \sum_{k=0}^{n} c_k z^k$ be a polynomial of degree $n \geq 1$ with complex coefficients. If $p(a) \neq 0$ for some $a \in \mathbb{C}$, then every disk D around $z = a$ contains an interior point b with $|p(b)| < |p(a)|$.

Proof. Without loss of generality, we may assume that $a = 0$ and $p(a) = 1$. If this is not the case, then we simply consider the polynomial $q(z) := \frac{p(z+a)}{p(a)}$, which satisfies $q(0)=1$.

Now, let us write $p(z) = 1 + c_1 z + \cdots + c_n z^n$ with m the smallest positive integer such that $c_m \neq 0$. We further define

$$r_1 := |c_m|^{-\frac{1}{m}} \qquad \text{and} \qquad r_2 := \begin{cases} \frac{|c_m|}{|c_{m+1}|+\cdots+|c_n|} & \text{if } m < n, \\ 1 & \text{if } m = n. \end{cases}$$

Let r be such that $0<r<\min\{1,r_1,r_2\}$ and let ζ be an m-th root of $-\frac{\overline{c_m}}{|c_m|}$ where $\overline{c_m}$ is the complex conjugate of c_m. Note that $|-\frac{\overline{c_m}}{|c_m|}|=1$. We claim that $b=r\zeta$ is the desired point as long as r is chosen to be small enough. There are two cases:

(i). If $m=n$, then $p(z)=1+c_nz^n$. Hence, $p(r\zeta)=1+c_n\cdot r^n\cdot\left(-\frac{\overline{c_n}}{|c_n|}\right)=1-|c_n|r^n\in\mathbb{R}$. Since $0<r<r_1$, we have $0<|c_n|r^n<1$, and thus $|p(r\zeta)|<1=|p(0)|$, as required.

(ii). If $m<n$, then we write $p(z)=1+c_mz^m+s(z)$. Note that $1+c_m(r\zeta)^m=1+c_m\cdot r^m\cdot\left(-\frac{\overline{c_m}}{|c_m|}\right)=1-|c_m|r^m\in\mathbb{R}$. Again, since $0<r<r_1$, we have $1-|c_m|r^m>0$. On the other hand, we deduce from $0<r<\min\{1,r_2\}$ and $|\zeta|=1$ that

$$\begin{aligned}|s(r\zeta)|&=\left|c_{m+1}r^{m+1}\zeta^{m+1}+\cdots+c_nr^n\zeta^n\right|\\&\le|c_{m+1}|r^{m+1}+\cdots+|c_n|r^n\\&\le(|c_{m+1}|+\cdots+|c_n|)r^{m+1}\\&<|c_m|r^m.\end{aligned}$$

We conclude that

$$\begin{aligned}|p(r\zeta)|&\le|1+c_m(r\zeta)^m|+|s(r\zeta)|\\&=1-|c_m|r^m+|s(r\zeta)|\\&<1=|p(0)|,\end{aligned}$$

and hence complete the proof. ■

R With the knowledge of complex analysis, the d'Alembert–Argand Lemma can be easily understood with recourse to the *maximum modulus principle*, asserting that if f is a holomorphic function, then the modulus $|f|$ cannot exhibit a strict local maximum that is properly within the domain of f. Now, if there exists an open disk D around $z=a$ such that for all $z\in D$, $|p(z)|\ge|p(a)|>0$, then $1/p(z)$ is holomorphic in D, and $|1/p(z)|$ reaches its local maximum at $a\in D$. But this violates the maximum modulus principle.

Now we are in a position to prove the Fundamental Theorem of Algebra.

Proof of Theorem 20.1. Let $p(z)$ be a complex polynomial of degree $n\ge1$. It suffices to show that $p(z)$ has at least one root z_1. Then we write $p(z)=(z-z_1)q(z)$ with $q(z)$ a polynomial of degree $n-1$, and apply induction on the degree.

To show that $p(z)$ has at least one root, we start by observing that $|p(z)|$ goes to infinity as $|z|$ goes to infinity. Hence, there exists some $R_1>0$ such that $|p(z)|>|p(0)|$ for all complex z with $|z|=R_1$. Now we apply Cauchy's minimum principle to the compact set $D:=\{z:|z|\le|R_1|\}$ and assume that $|p(z)|$ reaches the minimum at some $z=z_1$. Further, this z_1 is in the interior of D. If $p(z_1)\ne0$, then by the d'Alembert–Argand Lemma, we may find a certain $z_1'\in D$ such that $|p(z_1')|<|p(z_1)|$, thereby violating the assumption that $|p(z_1)|$ is the minimum. Hence, $p(z_1)=0$, which is our desired result. ■

20.2 Algebraic and transcendental numbers

Definition 20.1 An *algebraic number* is a complex number that is a root of a nonzero polynomial in one variable with integer (or, equivalently, rational) coefficients.

■ **Example 20.1** **(i).** Every rational number p/q is algebraic for it is the root of $qz-p$; **(ii).** Every m-th root of unity with m a positive integer is algebraic for it is a root of z^m-1; **(iii).** The number $\sqrt{2}+\sqrt{3}$ is algebraic for it is a root of z^4-10z^2+1. ■

Definition 20.2 Given an algebraic number, its *minimal polynomial* is a monic polynomial (i.e. polynomial with leading coefficient 1) with rational coefficients of least degree that has the number as a root. Further, this algebraic number is said to be of *degree n* if its minimal polynomial is of degree n.

Example 20.2 **(i).** Every rational number p/q is of degree 1 since its minimal polynomial is $z-\frac{p}{q}$; **(ii).** The number $e^{\frac{2\pi i}{3}} = \frac{-1+\sqrt{3}i}{2}$ is of degree 2 since its minimal polynomial is z^2+z+1. ■

Theorem 20.3 Given an algebraic number, its minimal polynomial is unique.

Proof. Assume that α is an algebraic number of degree n with two different minimal polynomials $p_1(z)$ and $p_2(z)$. Now, $p_1(z)$ and $p_2(z)$ are of degree n with $p_1(\alpha)=p_2(\alpha)=0$. Since $p_1(z)$ and $p_2(z)$ are monic, we know that the nonzero polynomial $q(z)=p_1(z)-p_2(z)$ with rational coefficients is of degree at most $n-1$. Noting that α is a root of $q(z)$, we are led to a contradiction as we have assumed that α is of degree n. ■

Theorem 20.4 Given an algebraic number α with minimal polynomial $p(z)$, if $f(z)$ is a polynomial with rational coefficients such that $f(\alpha)=0$, then we may write $f(z)=p(z)q(z)$ where $q(z)$ is also a polynomial with rational coefficients.

Proof. Noting that $p(z)$ is the minimal polynomial of α, and that $f(\alpha)=0$, the degree of $f(z)$ is no smaller than the degree of $p(z)$. Hence, we may write $f(z)=p(z)q(z)+r(z)$ using the Division Algorithm with the degree of $r(z)$ smaller than the degree of $p(z)$. Now we have $r(\alpha)=f(\alpha)-p(\alpha)q(\alpha)=0$. To ensure that $p(z)$ is the minimal polynomial of α, we must have that $r(z)$ is identical to zero, and hence that $f(z)=p(z)q(z)$. ■

Definition 20.3 Given an algebraic number α with minimal polynomial $p(z)$, its *algebraic conjugates*, or *conjugates* if there is no ambiguity, are the roots of $p(z)$. Normally, α itself is included in the set of conjugates of α.

Example 20.3 The number $e^{\frac{2\pi i}{3}} = \frac{-1+\sqrt{3}i}{2}$ has two conjugates $e^{\frac{2\pi i}{3}} = \frac{-1+\sqrt{3}i}{2}$ and $e^{-\frac{2\pi i}{3}} = \frac{-1-\sqrt{3}i}{2}$ since its minimal polynomial z^2+z+1 has the aforementioned two roots. ■

Our next object concerns the cardinality of the set of algebraic numbers.

Theorem 20.5 The set of algebraic numbers is countable.

Proof. Consider the set of non-constant polynomials with integer coefficients

$$\mathscr{P} := \{c_n z^n + \cdots + c_1 z + c_0 : c_0, \ldots, c_n \in \mathbb{Z},\ c_n \neq 0 \text{ and } n \geq 1\}.$$

For each $p(z)=c_n z^n+\cdots+c_1 z+c_0 \in \mathscr{P}$, we define $H(p) := n+|c_n|+\cdots+|c_0|$. Note that $H(p)\geq 2$. For every positive integer $N\geq 2$, there are finitely many polynomials p in $\mathscr{P}$ with $H(p)=N$. We label them as $p_{N,1}(z),\ldots,p_{N,k_N}(z)$. Now arranging these polynomials in the sequence

$$p_{2,1}(z), p_{2,2}(z), \ldots, p_{2,k_2}(z), p_{3,1}(z), p_{3,2}(z), \ldots, p_{3,k_3}(z), \ldots,$$

we are led to a one-to-one correspondence between polynomials in $\mathscr{P}$ and the set of natural numbers. But every algebraic number corresponds to at least one of these polynomials, and the number of algebraic numbers corresponding to any polynomial is finite by the Fundamental Theorem of Algebra. The claim therefore follows. ■

Definition 20.4 A complex number that is not algebraic is called *transcendental.*

Theorem 20.6 Almost all real and complex numbers are transcendental.

Proof. This is a direct consequence of Theorem 20.5 as both sets of real and complex numbers are uncountable. ■

Although almost all real and complex numbers are transcendental, it is in general not easy to determine if a given number is algebraic or transcendental. There are a few results that can be used to prove the transcendence of certain numbers. For example, the Lindemann–Weierstrass theorem implies that π, e^{α}, $\sin\alpha$, $\cos\alpha$, $\tan\alpha$, $\csc\alpha$, $\sec\alpha$, $\cot\alpha$ and their hyperbolic counterparts (with α algebraic and nonzero) are transcendental; the Gelfond–Schneider theorem implies that e^{π} and α^{β} (with α algebraic but not 0 or 1, and β irrational algebraic) are transcendental.

Note that every transcendental number is irrational, but the opposite is false. For instance, the number $\sqrt{2}$ is irrational but algebraic. However, it is unknown if the Apéry's constant $\zeta(3)$, which is irrational, is transcendental or not. Also, as pointed out in Sect. 19.2, the irrationality of $\pi+e$, πe, π/e, π^{e}, 2^{e}, $\log\pi$ or the Euler–Mascheroni constant γ is still mysterious, let alone their transcendence.

20.3 Transcendence of e

Our object in this section is the transcendence of e. The presented proof is attributed to the German mathematician David Hilbert (*Math. Ann.* **43** (1893), no. 2-3, 216–219).

Theorem 20.7 e is transcendental.

Proof. We argue by contradiction and assume that e is algebraic, of degree n. Then we may find a polynomial of degree n with integer coefficients such that

$$c_0+c_1e+\cdots+c_ne^n=0,$$

where c_0 and c_n are nonzero.

Now we define, with p a large prime such that $p>\max\{n,|c_0|\}$,

$$f(t):=\frac{t^{p-1}\big((t-1)\cdots(t-n)\big)^p}{(p-1)!}.$$

We further let

$$F(t):=\sum_{N\geq 0}f^{(N)}(t),$$

where $f^{(N)}$ is the N-th derivative of f. Note that this sum is indeed terminating. Repeatedly integrating by parts gives

$$\int_x^{\infty}f(t)e^{-t}dt=e^{-x}F(x).$$

Consider the two quantities:

$$S_1:=c_0\int_0^{\infty}f(t)e^{-t}dt+c_1e\int_1^{\infty}f(t)e^{-t}dt+\cdots+c_ne^n\int_n^{\infty}f(t)e^{-t}dt,$$

$$S_2:=c_1e\int_0^{1}f(t)e^{-t}dt+\cdots+c_ne^n\int_0^{n}f(t)e^{-t}dt.$$

Thus,

$$\begin{aligned}
S_1+S_2 &= c_0\int_0^\infty f(t)e^{-t}dt + c_1 e\int_0^\infty f(t)e^{-t}dt + \cdots + c_n e^n\int_0^\infty f(t)e^{-t}dt\\
&= (c_0+c_1e+\cdots+c_ne^n)\int_0^\infty f(t)e^{-t}dt\\
&= 0.
\end{aligned}$$

We first claim that S_1 is a nonzero integer. Note that

$$S_1 = c_0F(0)+c_1F(1)+\cdots+c_nF(n).$$

We shall show that $F(0)\in\mathbb{Z}\backslash p\mathbb{Z}$. To see this, we write $f(t)=g_0(t)h_0(t)$ where

$$g_0(t)=\frac{t^{p-1}}{(p-1)!} \qquad \text{and} \qquad h_0(t)=\big((t-1)\cdots(t-n)\big)^p.$$

Now,

$$f^{(N)}(0)=\sum_{i=0}^{N}\binom{N}{i}g_0^{(i)}(0)h_0^{(N-i)}(0)=\binom{N}{p-1}h_0^{(N-p+1)}(0)\in\mathbb{Z}.$$

Hence, $F(0)=\sum_{N\geq 0}f^{(N)}(0)\in\mathbb{Z}$. Further,

$$\begin{aligned}
F(0) &= \sum_{N\geq p-1}\binom{N}{p-1}h_0^{(N-p+1)}(0)\\
&\equiv h_0(0)=(-1)^{np}(n!)^p\not\equiv 0 \pmod p,
\end{aligned}$$

where we make use of the fact that $p>n$ is a prime. It follows that $F(0)\notin p\mathbb{Z}$, and hence that $F(0)\in\mathbb{Z}\backslash p\mathbb{Z}$. We shall also show that for $1\leq k\leq n$, $F(k)\in p\mathbb{Z}$. To see this, we write $f(t)=g_k(t)h_k(t)$ where

$$g_k(t)=\frac{(t-k)^p}{(p-1)!} \qquad \text{and} \qquad h_k(t)=\frac{t^{p-1}\big((t-1)\cdots(t-n)\big)^p}{(t-k)^p}.$$

Now,

$$f^{(N)}(k)=\sum_{i=0}^{N}\binom{N}{i}g_k^{(i)}(k)h_k^{(N-i)}(k)=p\binom{N}{p}h_k^{(N-p)}(k)\in p\mathbb{Z}.$$

Hence, $F(k)=\sum_{N\geq 0}f^{(N)}(k)\in p\mathbb{Z}$. At last, we have $p\nmid c_0$ since $p>|c_0|$. Consequently, $S_1\not\equiv 0 \pmod p$, which indicates that S_1 is a nonzero integer.

We next claim that $|S_2|\to 0$ as $p\to\infty$. To see this, it suffices to bound for each k with $0\leq k\leq n$,

$$\left|e^k\int_0^k f(t)e^{-t}dt\right|\leq e^n\int_0^n|f(t)|dt\leq e^n\cdot n\cdot\frac{n^{p-1}(n!)^p}{(p-1)!}\to 0,$$

as $p\to\infty$.

The above arguments imply that $S_1+S_2\neq 0$ for sufficiently large p, thereby violating the fact that $S_1+S_2=0$ as obtained earlier. Hence, e is not algebraic. ■

21. Number fields

21.1 Field extensions

In the previous lecture, we already experienced the basic background of algebraic numbers. Now we are going to look at a more abstract setting.

Definition 21.1 Let K and F be fields with $F \subset K$. Then F is called a *subfield* of K, and K is called an *extension field* of F. We usually write K/F, denoting that K is a *field extension* of F.

If a field K extends another field F, we can view K as a vector space over F. Therefore, we may determine a set of elements in K, say $B = \{b_i : i \in I\} \subset K$, where the set of indices is *not* necessarily countable, such that every element $y \in K$ can be **uniquely** written as $y = \sum_{i \in I} x_i b_i$ with $x_i \in F$ for all $i \in I$. This set B is called a *basis for K over F*.

Definition 21.2 The dimension of K as a vector space over F is called the *degree* of the extension K/F, denoted by $[K : F]$. In particular, $[K : F] = |B| = |I|$.

Proposition 21.1 Let $F \subset K$ be fields with $\{b_i : i \in I\}$ a basis for K over F. Then $\sum_{i \in I} x_i b_i = 0$ with $x_i \in F$ if and only if $x_i = 0$ for all i.

Proof. Note that $0 = \sum_{i \in I} 0 \cdot b_i$. The desired claim follows from the uniqueness of the representation. ■

Definition 21.3 We say K is a *finite extension* of F if $[K : F] < \infty$.

Example 21.1 **(i).** $\mathbb{C}$ is a finite extension of $\mathbb{R}$ with $[\mathbb{C} : \mathbb{R}] = 2$, and $\{1, i\}$ forms a basis. **(ii).** $\mathbb{C}$ is an extension of $\mathbb{Q}$ with $[\mathbb{C} : \mathbb{Q}] = \infty$, and the basis is uncountable. ■

Definition 21.4 If L, K and F are fields with $F \subset K \subset L$, we say K is an *intermediate field* of L and F.

Theorem 21.2 Let $F \subset K \subset L$ be fields. If the elements $\{\alpha_i \in K : i \in I\}$ form a basis for K over F and the elements $\{\beta_j \in L : j \in J\}$ form a basis for L over K, then the elements $\{\alpha_i \beta_j : i \in I,\ j \in J\}$ form a basis for L over F.

Proof. By definition, every $z \in L$ can be represented as $z = \sum_{j \in J} y_j \beta_j$ with $y_j \in K$. Further, each y_j can be represented as $y_j = \sum_{i\in I} x_{ij}\alpha_i$ with $x_{ij} \in F$. Hence, $z = \sum_{i\in I, j\in J} x_{ij}\alpha_i\beta_j$. Now we show the uniqueness of such representations, and it suffices to show that if $0 = \sum_{i\in I, j\in J} \tilde{x}_{ij}\alpha_i\beta_j$, then $\tilde{x}_{ij} = 0$ for all i and j. Noting that $0 = \sum_{j\in J} (\sum_{i\in I} \tilde{x}_{ij}\alpha_i)\beta_j$ and that β_j form a basis for L over K, we know from Proposition 21.1 that $\sum_{i\in I} \tilde{x}_{ij}\alpha_i = 0$ for all j. Using Proposition 21.1 again with the fact that α_i form a basis for K over F, we have $\tilde{x}_{ij} = 0$ for all i and j, as required. ■

Theorem 21.3 Let $F \subset K \subset L$ be fields. Then $[L:F] = [L:K][K:F]$.

Proof. In Theorem 21.2, we have $[K:F] = |I|$, $[L:K] = |J|$, and $[L:F] = |I||J|$. ■

Definition 21.5 *Number fields* are finite extensions of $\mathbb{Q}$.

21.2 Algebraicity

Definition 21.6 Let F be a field. We denote by $F[x]$ the set of polynomials in x whose coefficients are all in F, and call it the *ring of polynomials over* F. Further, we denote $F(x) = \left\{\frac{f(x)}{g(x)} : f, g \in F[x],\ g \neq 0\right\}$, and call it the *field of rational functions over* F.

Now we extend the concept of algebraicity discussed in Sect. 20.2.

NOTE Throughout, every field is assumed to be a subfield of $\mathbb{C}$ under the usual addition and multiplication, unless otherwise specified.

Definition 21.7 Let $F \subset \mathbb{C}$ be a field. We say $\alpha \in \mathbb{C}$ is *algebraic over* F if there is a nonzero polynomial $f(x) \in F[x]$ such that $f(\alpha) = 0$. Further, a polynomial in $F[x]$ is said to be a *minimal polynomial of* α *over* F if it is monic, has the lowest degree, and has α as its root. Finally, the *degree of* α *over* F is the degree of the minimal polynomial of α over F.

Definition 21.8 Let $F \subset \mathbb{C}$ be a field. We call K/F an *algebraic extension* if every element in K is algebraic over F.

We have considered some basic properties of algebraicity over $\mathbb{Q}$ in Sect. 20.2. Recall that in the proofs of Theorems 20.3 and 20.4, we used no specific properties of $\mathbb{Q}$ other than the fact that it is a field. Therefore, it is natural to transplant those arguments to the algebraicity over a generic subfield of $\mathbb{C}$.

Theorem 21.4 Let α be algebraic over $F \subset \mathbb{C}$. Then its minimal polynomial is unique.

Theorem 21.5 Let α be algebraic over $F \subset \mathbb{C}$ with minimal polynomial $p(x)$. If $f(x) \in F[x]$ is such that $f(\alpha) = 0$, then we may write $f(x) = p(x)q(x)$ where $q(x) \in F[x]$.

The next result indicates that algebraic elements over a subfield of $\mathbb{C}$ are not rare.

Theorem 21.6 Let $F \subset \mathbb{C}$ be a field and let K/F be a finite extension. Then every element in K is algebraic over F, of degree at most $[K:F]$.

Proof. Let $[K:F] = n < \infty$. For any $\alpha \in K$, we see that $1, \alpha, \alpha^2, \dots, \alpha^n$ are $n+1$ elements in K. Hence, they are linearly dependent over F, meaning that there is a nontrivial linear combination of $1, \alpha, \alpha^2, \dots, \alpha^n$ of value 0 such that the coefficients are in F. This gives a nonzero polynomial of degree at most n in $F[x]$ having α as a root. ■

21.3 Algebraic conjugates

Definition 21.9 Let F be a field. A polynomial $f \in F[x]$ is said to be *irreducible over* F if and only if whenever $f = gh$ with $g, h \in F[x]$, we have g or h constant.

Theorem 21.7 Let $F \subset \mathbb{C}$ be a field. For every algebraic α over F, a monic polynomial $f(x) \in F[x]$ such that $f(\alpha) = 0$ is the minimal polynomial of α if and only if $f(x)$ is irreducible over F.

Proof. We start with necessity. If $f(x)$ is not irreducible over F, then we may write $f(x) = s(x)t(x)$ with $s, t \in F[x]$ and $1 \leq \deg s, \deg t < \deg p$. Further, we have $s(\alpha)t(\alpha) = f(\alpha) = 0$. Then either $s(\alpha) = 0$ or $t(\alpha) = 0$, thereby violating the assumption that $f(x)$ is the minimal polynomial of α.

We then prove sufficiency. Assume that the minimal polynomial of α is $p(x)$. Then by Theorem 21.5, there exists a polynomial $q \in F[x]$ such that $f(x) = p(x)q(x)$. However, since the monic polynomial $f(x)$ is irreducible, we must have $q(x) = 1$, and hence $p(x) = f(x)$. ■

Theorem 21.8 Let $F \subset \mathbb{C}$ be a field. If $f \in F[x]$ of degree n is irreducible over F, then f has n distinct roots in $\mathbb{C}$.

Proof. By the Fundamental Theorem of Algebra, we know that f has n roots in $\mathbb{C}$. Now we show that these roots are distinct. Note that if $n = 1$, the statement is trivial. So we assume that $n \geq 2$. Suppose that there is a root α with multiplicity at least two. Then as a polynomial over $\mathbb{C}$, we may write $f(x) = (x - \alpha)^2 g(x)$ with $g \in \mathbb{C}[x]$. Taking the derivative $f'(x) = 2(x - \alpha)g(x) + (x - \alpha)^2 g'(x)$, we observe that $f'(\alpha) = 0$. If the minimal polynomial of α is $p(x)$, then $1 \leq \deg p \leq n - 1$ since f' is in $F[x]$ and $\deg f' = \deg f - 1 = n - 1$. Recalling that $f(\alpha) = 0$, we conclude from Theorem 21.5 that $f(x) = p(x)q(x)$ with $q(x) \in F[x]$. However, by the fact that $1 \leq \deg p \leq n - 1$ so that $1 \leq \deg q \leq n - 1$, we find that f is not irreducible, thereby leading to a contradiction. ■

Corollary 21.9 Let $F \subset \mathbb{C}$ be a field. For every algebraic α over F of degree n, its minimal polynomial $p(x)$ has n distinct roots in $\mathbb{C}$.

Proof. This is a direct implication of Theorems 21.7 and 21.8. ■

Now we extend the concept of conjugacy in Definition 20.3.

Definition 21.10 Let α be algebraic over $F \subset \mathbb{C}$ with minimal polynomial $p(x)$ of degree n. Let $\alpha = \alpha_1, \alpha_2, \ldots, \alpha_n$ be the (distinct) roots of $p(x)$. We call $\alpha_1, \alpha_2, \ldots, \alpha_n$ the *algebraic conjugates* or *conjugates of* α *over* F.

Corollary 21.10 If α is algebraic over $F \subset \mathbb{C}$ with minimal polynomial $p(x)$, then so is every conjugate $\tilde{\alpha}$ of α over F. Consequently, the degree of $\tilde{\alpha}$ over F equals the degree of α over F.

Proof. By Theorem 21.7, $p(x)$ is irreducible over F. Also, it is known that $p(\tilde{\alpha}) = 0$. Hence, with Theorem 21.7 applied again, the minimal polynomial of $\tilde{\alpha}$ over F is also $p(x)$. ■

21.4 $F[\alpha]$ vs $F(\alpha)$

Theorem 21.11 Let $F \subset \mathbb{C}$ be a field. Let α be algebraic over F of degree n. The elements $\{1, \alpha, \dots, \alpha^{n-1}\}$ form a basis for $F[\alpha]$ over F.

Proof. First, we know that $\{1, \alpha, \dots, \alpha^{n-1}\}$ are linearly independent over F. Otherwise, if there exist $a_i \in F$ $(0 \le i \le n-1)$, not all zero, such that $\sum_{i=0}^{n-1} a_i \alpha^i = 0$, then we are led to a contradiction as the minimal polynomial of α is of degree n. Now we show that every element in $F[\alpha]$ can be written as a linear combination of $1, \alpha, \dots, \alpha^{n-1}$. We know that every element in $F[\alpha]$ is of the form $f(\alpha)$ where $f(x) \in F[x]$. Assuming that the minimal polynomial of α is $p(x)$, we then write $f(x) = p(x)q(x) + r(x)$ with $q, r \in F[x]$ and $\deg r < \deg p = n$. It follows that $f(\alpha) = p(\alpha)q(\alpha) + r(\alpha) = r(\alpha)$. Finally, since $\deg r < n$, we find that $r(\alpha)$ is a linear combination of $1, \alpha, \dots, \alpha^{n-1}$, as required. ■

Before moving forward, let us briefly consider the divisibility for polynomials over F, as an analog to that for integers discussed in Lecture 2.

Definition 21.11 Let $f, g \in F[x]$. We say that g *divides* f, denoted by $g \mid f$, if there exists a polynomial $h \in F[x]$ such that $f(x) = g(x)h(x)$.

Definition 21.12 Let $f, g \in F[x]$, not both zero. There exists a unique polynomial $d \in F[x]$, up to multiplying a nonzero constant in F, such that d divides both f and g, and such that if $\delta \in F[x]$ divides f and g, then $\delta \mid d$. This polynomial $d(x)$ is called the *greatest common divisor* of $f(x)$ and $g(x)$, denoted by $d(x) = (f(x), g(x))$.

To get this polynomial $d(x)$, we may still use the Euclidean Algorithm, but this time for polynomials over F. Without loss of generality, we assume that $\deg f \ge \deg g$ and $g \ne 0$. We also put $r_{-1}(x) = f(x)$ and $r_0(x) = g(x)$. Now we iteratively write

$$\begin{aligned}
r_{-1}(x) &= q_1(x) r_0(x) + r_1(x), && \deg r_1 < \deg r_0;\\
r_0(x) &= q_2(x) r_1(x) + r_2(x), && \deg r_2 < \deg r_1;\\
&\cdots\\
r_{k-2}(x) &= q_k(x) r_{k-1}(x) + r_k(x), && \deg r_k < \deg r_{k-1};\\
r_{k-1}(x) &= q_{k+1}(x) r_k(x) + 0.
\end{aligned}$$

Then $d(x) = r_k(x)$ is as required.

We are also able to establish a Bézout-type identity parallel to that in Theorem 2.5.

Theorem 21.12 (Bézout's Identity for Polynomials). Let $f, g \in F[x]$, not both zero, and denote $d(x) = (f(x), g(x))$. Then there exist polynomials $u, v \in F[x]$ such that $d(x) = f(x)u(x) + g(x)v(x)$.

Proof. We only need the fact that the set $S = \{f(x)u(x) + g(x)v(x) : u, v \in F[x]\}$ is closed under addition and scalar multiplication (of polynomials over F). From the above Euclidean Algorithm, we iteratively have $r_1(x) \in S$, $r_2(x) \in S$, $\dots$, and finally, $d(x) = r_k(x) \in S$. ■

Theorem 21.13 Let $F \subset \mathbb{C}$ be a field. Let α be algebraic over F. We have $F(\alpha) = F[\alpha]$. In particular, if the degree of α over F is n, then $F(\alpha) = F[\alpha]$ is a finite extension of F of degree n.

Proof. It is plain that $F[\alpha] \subset F(\alpha)$. Hence, we only need to show that $F(\alpha) \subset F[\alpha]$. Let $\theta \in F(\alpha)$. Then $\theta = \frac{f(\alpha)}{g(\alpha)}$ with $f, g \in F[x]$. Note that $g(\alpha) \ne 0$. Let p be the minimal polynomial of α over F. By Theorem 21.7, p is irreducible over F. Also, we have that $p \nmid g$; if not, then

$g(x) = p(x)q(x)$ with $q \in F[x]$, and hence $g(\alpha) = p(\alpha)q(\alpha) = 0$, which gives a contradiction. Hence, $(p(x), g(x)) = 1$, and by Theorem 21.12, there are polynomials $u, v \in F[x]$ such that $p(x)u(x) + g(x)v(x) = 1$. Now, $g(\alpha)v(\alpha) = 1$, implying that $\theta = f(\alpha)v(\alpha) \in F[\alpha]$, as proposed.

Now noting that $F(\alpha)$ is a field containing F, we have that $F(\alpha) = F[\alpha]$ is an extension of F. To show that this extension is of degree n, we only need to recall from Theorem 21.11 that the dimension of $F[\alpha]$ as a vector space over F is n. ■

Corollary 21.14 Let $F \subset \mathbb{C}$ be a field. Let $\alpha, \beta, \dots, \gamma$ be algebraic over F. Then we have $F(\alpha, \beta, \dots, \gamma) = F[\alpha, \beta, \dots, \gamma]$. In particular, if the degrees of $\alpha, \beta, \dots, \gamma$ over F are $n_\alpha, n_\beta, \dots, n_\gamma$, respectively, then $F(\alpha, \beta, \dots, \gamma) = F[\alpha, \beta, \dots, \gamma]$ is a finite extension of F of degree at most $n_\alpha n_\beta \cdots n_\gamma$.

Proof. The first part follows from an easy exercise of induction. Let $K = F(\beta, \dots, \gamma)$, which is an extension field of F. Note that $F(\alpha, \beta, \dots, \gamma) = K(\alpha) = K[\alpha]$ by Theorem 21.13 as α is algebraic over F, and hence also over K. Now, by the inductive assumption, we have $K = F[\beta, \dots, \gamma]$, and hence $K[\alpha] = F[\beta, \dots, \gamma][\alpha] = F[\alpha, \beta, \dots, \gamma]$.

For the second part, let $p(x)$ be the minimal polynomial of α over F. Then $p(x)$ is also over K. Therefore, the degree of α over K is at most $\deg p = n_\alpha$. By Theorem 21.13,

$$[F(\alpha, \beta, \dots, \gamma) : K] = [K(\alpha) : K] \le n_\alpha.$$

Finally, we have

$$[F(\alpha, \beta, \dots, \gamma) : F] = [F(\alpha, \beta, \dots, \gamma) : K][K : F] \le n_\alpha [K : F].$$

It is concluded that $[F(\alpha, \beta, \dots, \gamma) : F] \le n_\alpha n_\beta \cdots n_\gamma$ by induction. ■

We are then led to the transitivity of algebraicity of field extensions.

Theorem 21.15 (Transitivity of Algebraicity). Let $F \subset K \subset L$ be three subfields of $\mathbb{C}$. If L is algebraic over K and K is algebraic over F, then L is algebraic over F.

Proof. Let $\alpha \in L$. Since α is algebraic over K, it is a root of $a_n x^n + \cdots + a_0$ with $a_i \in K$. Further, each a_i is algebraic over F, and by Corollary 21.14, $M := F(a_0, \dots, a_n)$ is a finite extension of F. On the other hand, α is also algebraic over the field M. Hence, $F(a_0, \dots, a_n, \alpha) = M(\alpha)$ is a finite extension of M. It follows that $F(a_0, \dots, a_n, \alpha)$ is a finite extension of F as

$$[F(a_0, \dots, a_n, \alpha) : F] = [F(a_0, \dots, a_n, \alpha) : M][M : F].$$

Finally, by Theorem 21.6, $\alpha \in F(a_0, \dots, a_n, \alpha)$ is algebraic over F. ■

21.5 Field of algebraic elements

Now we are ready to claim the structure of algebraic elements over a subfield of $\mathbb{C}$.

Theorem 21.16 Let $F \subset \mathbb{C}$ be a field. Let α and β be algebraic over F. Then $\alpha + \beta$ and $\alpha\beta$ are algebraic over F. In particular, algebraic elements over F form a field.

Proof. Let $K = F[\alpha]$. Then $\alpha + \beta$ and $\alpha\beta$ are in $K[\beta]$. We know from Theorem 21.13 that $K = F[\alpha]$ is a finite extension of F, and that $K[\beta]$ is a finite extension of K. Now by

Theorem 21.3, $K[\beta]$ is a finite extension of F, and hence all elements in $K[\beta]$, including $\alpha+\beta$ and $\alpha\beta$, are algebraic over F with recourse to Theorem 21.6.

To show that algebraic elements over F form a field, it remains to prove that the additive and multiplicative inverses of any nonzero algebraic element are algebraic. Let $\alpha \neq 0$ be algebraic over F with minimal polynomial $p(x) = a_n x^n + \cdots + a_1 x + a_0$. Then $a_n \alpha^n + \cdots + a_1 \alpha + a_0 = 0$, and hence $(-1)^n a_n (-\alpha)^n + \cdots - a_1(-\alpha) + a_0 = 0$ and $a_n + \cdots + a_1(\alpha^{-1})^{n-1} + a_0(\alpha^{-1})^n = 0$, yielding the algebraicity of $-\alpha$ and α^{-1}. ■

The above proof of the algebraicity of $\alpha+\beta$ and $\alpha\beta$ over F is not constructive at all. A natural desire is finding polynomials over F having $\alpha+\beta$ or $\alpha\beta$ as a root. For this purpose, we need some knowledge of matrix theory, namely the *Kronecker product* of matrices, named after the German mathematician Leopold Kronecker.

Definition 21.13 Let F be a field. The *Kronecker product* of an $m \times n$ matrix $A = (a_{ij}) \in \mathrm{Mat}_{m,n}(F)$ and a $p \times q$ matrix $B = (b_{ij}) \in \mathrm{Mat}_{p,q}(F)$ is the $mp \times nq$ block matrix

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix} \in \mathrm{Mat}_{mp,nq}(F).$$

Proposition 21.17 Let A, B, C, D be matrices, $\mathbf{0}$ be the zero column vector, and k be a scalar. Then

(i) $A \otimes (B+C) = A \otimes B + A \otimes C$;
(ii) $(B+C) \otimes A = B \otimes A + C \otimes A$;
(iii) $(kA) \otimes B = A \otimes (kB) = k(A \otimes B)$;
(iv) $A \otimes \mathbf{0} = \mathbf{0} \otimes A = \mathbf{0}$;
(v) $A \otimes (B \otimes C) = (A \otimes B) \otimes C$;
(vi) $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$.

Proof. All these relations can be examined directly by the definition of Kronecker product. In particular, we show (vi). First, it follows from $A = (a_{ih})$ and $C = (c_{hj})$ that $A \otimes B = (a_{ih}B)$ and $C \otimes D = (c_{hj}D)$. Hence, the (i,j)-th block of $(A \otimes B)(C \otimes D)$ is

$$\sum_{h=1}^{n} (a_{ih}B)(c_{hj}D) = \left(\sum_{h=1}^{n} a_{ih}c_{hj} \right) BD.$$

On the other hand, the (i,j)-th entry of AC is $\sum_{h=1}^{n} a_{ih}c_{hj}$, thereby implying that the (i,j)-th block of $(AC) \otimes (BD)$ is also $\left(\sum_{h=1}^{n} a_{ih}c_{hj}\right) BD$. ■

Now let

$$f(x) = c_0 + c_1 x + \cdots + c_{n-1}x^{n-1} + x^n$$

be a monic polynomial over F. Its *companion matrix* is given by

$$C(f) = \begin{pmatrix} 0 & 0 & \cdots & 0 & -c_0 \\ 1 & 0 & \cdots & 0 & -c_1 \\ 0 & 1 & \cdots & 0 & -c_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -c_{n-1} \end{pmatrix}.$$

It is plain to verify that the characteristic polynomial of $C(f)$ is $f(x)$.

Lemma 21.18 Let A be an $m\times m$ matrix and B an $n\times n$ matrix. If u is an eigenvector of A for eigenvalue α and v is an eigenvector of B for eigenvalue β, then $u\otimes v$ is an eigenvector of $A\otimes B$ for eigenvalue $\alpha\beta$, and $u\otimes v$ is an eigenvector of $A\otimes I_n+I_m\otimes B$ for eigenvalue $\alpha+\beta$, where I_k is the $k\times k$ identity matrix.

Proof. Note that $Au=\alpha u$ and $Bv=\beta v$. For the first part, we have

$$(A\otimes B)(u\otimes v)=(Au)\otimes(Bv)=(\alpha u)\otimes(\beta v)=(\alpha\beta)(u\otimes v).$$

Also,

$$\begin{aligned}(A\otimes I_n+I_m\otimes B)(u\otimes v)&=(A\otimes I_n)(u\otimes v)+(I_m\otimes B)(u\otimes v)\\&=(Au)\otimes(I_nv)+(I_mu)\otimes(Bv)\\&=(\alpha u)\otimes v+u\otimes(\beta v)\\&=(\alpha+\beta)(u\otimes v),\end{aligned}$$

confirming the second part. ■

Theorem 21.19 Let $F\subset\mathbb{C}$ be a field. Let α and β be algebraic over F, of degree m and n, respectively. For $f(x)$ and $g(x)$ monic polynomials over F such that $f(\alpha)=0$ and $g(\beta)=0$, we denote by A and B the companion matrix of $f(x)$ and $g(x)$, respectively. Then the characteristic polynomial $s(x)$ (resp. $p(x)$) of $A\otimes I_n+I_m\otimes B$ (resp. $A\otimes B$) is monic over F such that $s(\alpha+\beta)=0$ (resp. $p(\alpha\beta)=0$).

Proof. The claim that $s(x)$ and $p(x)$ are monic over F is plain since A and B are over the field F, so are $A\otimes I_n+I_m\otimes B$ and $A\otimes B$. Further, $f(\alpha)=0$ and $g(\beta)=0$ imply that α is an eigenvalue of A and that β is an eigenvalue of B. We thus obtain $s(\alpha+\beta)=0$ and $p(\alpha\beta)=0$ from Lemma 21.18. ■

Example 21.2 Let $\alpha=\sqrt{2}$ and $\beta=\sqrt[3]{3}$. Then $f(x)=x^2-2$ and $g(x)=x^3-3$, and hence,

$$A=\begin{pmatrix}0&2\\1&0\end{pmatrix}\qquad\text{and}\qquad B=\begin{pmatrix}0&0&3\\1&0&0\\0&1&0\end{pmatrix}.$$

We further compute that

$$A\otimes I_3+I_2\otimes B=\begin{pmatrix}0&0&3&2&0&0\\1&0&0&0&2&0\\0&1&0&0&0&2\\1&0&0&0&0&3\\0&1&0&1&0&0\\0&0&1&0&1&0\end{pmatrix}\qquad\text{and}\qquad A\otimes B=\begin{pmatrix}0&0&0&0&0&6\\0&0&0&2&0&0\\0&0&0&0&2&0\\0&0&3&0&0&0\\1&0&0&0&0&0\\0&1&0&0&0&0\end{pmatrix},$$

and hence that $s(x)=x^6-6x^4-6x^3+12x^2-36x+1$ and $p(x)=x^6-72$. Finally, it is easy to verify that $s(\sqrt{2}+\sqrt[3]{3})=0$ and $p(\sqrt{2}\cdot\sqrt[3]{3})=0$. ■

The above construction can be found at a *Stackexchange* post of Robert Israel:

https://math.stackexchange.com/q/1283161

Israel attributed this folklore to Dave Boyd, Olga Taussky, and Richard Dedekind, in chronological order.

22. Embeddings

22.1 Embeddings

Definition 22.1 Let R and S be fields (resp. rings). A *field* (resp. *ring*) homomorphism is a map $\sigma : R \to S$ such that for all $a, b \in R$, the map σ is

(i) *addition preserving*: $\sigma(a+b) = \sigma(a) + \sigma(b)$;
(ii) *multiplication preserving*: $\sigma(ab) = \sigma(a)\sigma(b)$;
(iii) *multiplicative identity preserving*: $\sigma(1_R) = 1_S$.

R There are several basic facts about field or ring homomorphisms:

(i) $\sigma(0_R) = 0_S$. This is because $\sigma(0_R) = \sigma(0_R + 0_R) = \sigma(0_R) + \sigma(0_R)$.
(ii) $\sigma(-a) = -\sigma(a)$. This is because $\sigma(-a) + \sigma(a) = \sigma(-a+a) = \sigma(0_R) = 0_S$.
(iii) *Every field homomorphism is injective. In particular, we have $\sigma(a) = 0_S$ if and only if $a = 0_R$.* Otherwise suppose that there are two distinct $a, b \in R$ such that $\sigma(a) = \sigma(b)$. If we put $c = a - b \neq 0_R$, then $\sigma(c) = \sigma(a-b) = \sigma(a) - \sigma(b) = 0_S$. Now, $1_S = \sigma(1_R) = \sigma(cc^{-1}) = \sigma(c)\sigma(c^{-1}) = 0_S$. However, in the field S, we have required that the additive identity 0_S and the multiplicative identity 1_S are different.

Lemma 22.1 Let $F \subset \mathbb{C}$ be a field and let K/F be an extension. Let $\alpha \in K$ be algebraic over F of degree n with minimal polynomial $x^n + a_{n-1}x^{n-1} + \cdots + a_0$. For σ a field homomorphism of K in $\mathbb{C}$, $\sigma(\alpha)$ is a root of $x^n + \sigma(a_{n-1})x^{n-1} + \cdots + \sigma(a_0)$.

Proof. We have

$$\begin{aligned} 0 = \sigma(0) &= \sigma\left(\alpha^n + a_{n-1}\alpha^{n-1} + \cdots + a_0\right) \\ &= \sigma(\alpha)^n + \sigma(a_{n-1})\sigma(\alpha)^{n-1} + \cdots + \sigma(a_0), \end{aligned}$$

as required. ■

Definition 22.2 Let R and S be fields (resp. rings). An injective field (resp. ring) homomorphism $\sigma : R \to S$ is said to be an *embedding* of R in S.

Theorem 22.2 Let $F \subset \mathbb{C}$ be a field. Let α be algebraic over F of degree n with minimal polynomial $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0 \in F[x]$. Then every embedding σ of F in $\mathbb{C}$ extends to exactly n distinct embeddings $\sigma_1, \ldots, \sigma_n$ of $F(\alpha)$ in $\mathbb{C}$, and each σ_i sends α to one of the n distinct roots of $q(x) = x^n + \sigma(a_{n-1})x^{n-1} + \cdots + \sigma(a_0)$ in $\mathbb{C}$.

In particular, if α has conjugates $\alpha_1, \ldots, \alpha_n$ over F, then there are exactly n embeddings $\tau_1, \ldots, \tau_n$ of $F(\alpha)$ in $\mathbb{C}$ that fix F, and each τ_i sends α to α_i.

Proof. Note that by Theorem 21.7, $p(x)$ is irreducible over F, and hence $q(x)$ is irreducible over the field $\sigma(F)$. It follows from Theorem 21.8 that $q(x)$ has n distinct roots $\beta_1, \ldots, \beta_n$ in $\mathbb{C}$. For each β_i, we get a field homomorphism σ_i by sending α to β_i, and hence,

$$\begin{array}{rccc} \sigma_i : & F(\alpha) = F[\alpha] & \to & \mathbb{C} \\ & \sum_j c_j \alpha^j & \mapsto & \sum_j \sigma(c_j)\beta_i^j \end{array}$$

We have obtained n distinct extensions of σ to embeddings of $F(\alpha)$ in $\mathbb{C}$. Further, there are no other such embeddings by Lemma 22.1, which tells us that α must be sent to a root of $q(x)$.

For the second part, we take σ to be the trivial embedding $\mathrm{id} : F \to \mathbb{C}$ with $\mathrm{id}(x) = x$ for every $x \in F$. Then for each coefficient $a_i \in F$ of $p(x)$, we have $\sigma(a_i) = \mathrm{id}(a_i) = a_i$, and hence $q(x) = p(x)$. Now, the roots of $q(x)$ in $\mathbb{C}$ are the conjugates $\alpha_1, \ldots, \alpha_n$ of α over F. ■

Theorem 22.3 Let $F \subset \mathbb{C}$ be a field and let K be a finite extension of F. Then every embedding σ of F in $\mathbb{C}$ extends to exactly $[K : F]$ distinct embeddings of K in $\mathbb{C}$.

Proof. We apply induction on $[K : F] = N$. The statement is trivial if $N = 1$. Now we assume that $N \geq 2$. Let $\alpha = K \backslash F$ and assume that the degree of α over F is n. Note that $n > 1$ since $\alpha \notin F$. We know from Theorem 22.2 that there are exactly n distinct extensions $\sigma_1, \ldots, \sigma_n$ of σ to embeddings of $F(\alpha)$ in $\mathbb{C}$. Further, since $[K : F(\alpha)] = \frac{N}{n} < N$ as $n > 1$, we derive from the inductive hypothesis that each σ_i has exactly $\frac{N}{n}$ distinct extensions to embeddings of K in $\mathbb{C}$ and so we have exactly $n \times \frac{N}{n} = N$ distinct embeddings of K in $\mathbb{C}$ extending σ. ■

Definition 22.3 Let $F \subset \mathbb{C}$ be a field and let K be a finite extension of F. If σ is an embedding of K in $\mathbb{C}$ that fixes F, i.e. σ is extended by the embedding $\mathrm{id} : F \to \mathbb{C}$ with $\mathrm{id}(x) = x$ for every $x \in F$, then we say σ is an *embedding of K in $\mathbb{C}$ over F*.

Corollary 22.4 Let $F \subset \mathbb{C}$ be a field and let K be a finite extension of F. Then there are exactly $[K : F]$ distinct embeddings of K in $\mathbb{C}$ over F.

Proof. In Theorem 22.3, we take $\sigma = \mathrm{id}$. ■

22.2 Finite extensions are simple

Now we are in a position to show that every finite extension of $F \subset \mathbb{C}$ is *simple*, that is, it is generated by the adjunction of a single element.

Theorem 22.5 Let $F \subset \mathbb{C}$ be a field and let K/F be a finite extension. Then there is an element $\theta \in K$ such that $K = F(\theta)$.

Proof. Since K/F is a finite extension, there exist algebraic elements $\alpha_1, \ldots, \alpha_n$ over F such that $K = F(\alpha_1, \ldots, \alpha_n)$. By induction, it suffices to show the case where $n = 2$.

Suppose that $K = F(\alpha,\beta)$ with α,β algebraic over F. Let $\alpha = \alpha_1,\alpha_2,\dots,\alpha_r$ be the conjugates of α over F and $\beta = \beta_1,\beta_2,\dots,\beta_s$ be the conjugates of β over F. Note that F is infinite since $\mathbb{Q} \subset F$. We may choose some $c \in F$ such that $c \neq -\frac{\alpha_1-\alpha_i}{\beta_1-\beta_j}$ for all $1 \le i \le r$ and $2 \le j \le s$. Put $\theta = \alpha + c\beta = \alpha_1 + c\beta_1$. It is trivial that $F(\theta) \subset F(\alpha,\beta)$.

Now we show that $F(\alpha,\beta) \subset F(\theta)$ and hence obtain the desired θ. Note that it suffices to prove that $\alpha,\beta \in F(\theta)$. Further, since $\alpha = \theta - c\beta$, we only need to prove that $\beta \in F(\theta)$.

Let $p(x)$ be the minimal polynomial of α over F and let $q(x)$ be the minimal polynomial of β over F. Considering polynomials $p(\theta - cx)$ and $q(x)$, we observe that $x = \beta = \beta_1$ is a common root of them. If there is a common root $x_0 \in \mathbb{C}$ other than $\beta = \beta_1$, then it must be $x_0 = \beta_k$ for some $2 \le k \le s$ since $q(x_0) = 0$. Further, $p(\theta - cx_0) = p(\theta - c\beta_j) = 0$ implies that $\theta - c\beta_j = \alpha_i$ for some $1 \le i \le r$. Then $c = -\frac{\alpha_1-\alpha_i}{\beta_1-\beta_j}$, thereby violating our choice of c. Hence, $x = \beta = \beta_1$ is the only common root of $p(\theta - cx)$ and $q(x)$. By Corollary 21.9, the roots of $q(x)$ are distinct. Hence, $(p(\theta - cx), q(x)) = x - \beta$ over $\mathbb{C}$.

On the other hand, we find that the polynomials $p(\theta - cx)$ and $q(x)$ are in $F(\theta)[x]$. Let $f(x)$ be the minimal polynomial of β over $F(\theta)$. Then by Theorem 21.5, $f(x)$ divides $p(\theta - cx)$ and $q(x)$ over $F(\theta)$, and thus over $\mathbb{C}$.

Finally, we know from Definition 21.12 that $f(x) \mid (p(\theta - cx), q(x)) = (x - \beta)$ over $\mathbb{C}$. Since $\deg f \ge 1$ and $f(x)$ is monic, it follows that $f(x) = x - \beta$. However, $f(x) \in F(\theta)[x]$, thereby indicating that we must have $\beta \in F(\theta)$. ■

Example 22.1 Consider $\mathbb{Q}(\sqrt{2},\sqrt[3]{3})$. We find that the minimal polynomial of $\sqrt{2}$ over $\mathbb{Q}$ is $x^2 - 2$, and hence its conjugates are $\sqrt{2}$ and $-\sqrt{2}$. Similarly, the minimal polynomial of $\sqrt[3]{3}$ over $\mathbb{Q}$ is $x^3 - 3$, and its conjugates are $\sqrt[3]{3}$, $\omega\sqrt[3]{3}$ and $\omega^2\sqrt[3]{3}$ where $\omega = e^{2\pi i/3}$. It is plain that 1 is not equal to any of $-\frac{\sqrt{2}-(\pm\sqrt{2})}{\sqrt[3]{3}-(\omega\sqrt[3]{3})}$ and $-\frac{\sqrt{2}-(\pm\sqrt{2})}{\sqrt[3]{3}-(\omega^2\sqrt[3]{3})}$. Thus, we have $\mathbb{Q}(\sqrt{2},\sqrt[3]{3}) = \mathbb{Q}(\sqrt{2}+\sqrt[3]{3})$. ■

22.3 Automorphisms of a field extension

Definition 22.4 Let R and S be fields (resp. rings). A homomorphism $\sigma : R \to S$ is called an *isomorphism* if it is bijective. In addition, an isomorphism from a field or a ring to itself is called an *automorphism*.

Definition 22.5 Let K be an algebraic extension of a field $F \subset \mathbb{C}$. Denote by $\mathrm{Aut}(K/F)$ the set of automorphisms $\sigma : K \to K$ leaving elements in F fixed, namely, $\sigma(x) = x$ whenever $x \in F$.

R $\mathrm{Aut}(K/F)$ forms a group under composition.

Proposition 22.6 Let $F \subset \mathbb{C}$ be a field and let K/F be an algebraic extension. Let $\alpha \in K$ with minimal polynomial $p(x)$ over F. Then for any $\sigma \in \mathrm{Aut}(K/F)$, $\sigma(\alpha)$ is a root of $p(x)$ in K.

Proof. This is a direct application of Lemma 22.1 with the fact that σ fixes F, and hence all coefficients of $p(x)$. ■

Theorem 22.7 Let $F \subset \mathbb{C}$ be a field and let K/F be a finite extension. We have $|\mathrm{Aut}(K/F)| \le [K : F]$.

Proof. We know from Theorem 22.5 that there is an element $\theta \in K$ such that $K = F(\theta)$. Also, by Theorem 21.13, the degree of θ over F is $[F(\theta) : F] = [K : F] =: n$. Since $K = F(\theta) =$

$F[\theta]$, we may write every $x \in K$ as $x = \sum_i c_i\theta^i$ with $c_i \in F$, and hence for $\sigma \in \mathrm{Aut}(K/F)$, the value of $\sigma(x) = \sum_i c_i\sigma(\theta)^i$ is uniquely determined by $\sigma(\theta)$. Finally, Proposition 22.6 asserts that there are at most n possibilities of $\sigma(\theta)$. Here, we shall use "at most" as there might be roots of the minimal polynomial of θ over F falling outside of K. ■

Theorem 22.8 Let $F \subset \mathbb{C}$ be a field. Let α be algebraic over F. If σ is an embedding of $F(\alpha)$ in $\mathbb{C}$ over F such that $\sigma(\alpha) \in F(\alpha)$, then $\sigma \in \mathrm{Aut}(F(\alpha)/F)$.

Proof. Let $p(x)$ be the minimal polynomial of α over F. Lemma 22.1 tells us that $\sigma(\alpha)$ is also a root of $p(x)$ since σ fixes F. Hence, by Corollary 21.10, $\sigma(\alpha)$ also has minimal polynomial $p(x)$ over F. It is then a consequence of Theorem 21.13 that $[F(\sigma(\alpha)) : F] = \deg p = [F(\alpha) : F]$. Since $\sigma(\alpha) \in F(\alpha)$, we have $F(\sigma(\alpha)) \subset F(\alpha)$. Thus,

$$[F(\alpha) : F(\sigma(\alpha))] = \frac{[F(\alpha) : F]}{[F(\sigma(\alpha)) : F]} = 1,$$

yielding that $F(\sigma(\alpha)) = F(\alpha)$.

To show that $\sigma \in \mathrm{Aut}(F(\alpha)/F)$, it is sufficient to prove that σ is onto $F(\alpha)$. Note that every $x \in F(\alpha) = F(\sigma(\alpha)) = F[\sigma(\alpha)]$ can be written as $x = \sum_i c_i\sigma(\alpha)^i$ with $c_i \in F$. It turns out that

$$x = \sum_i c_i\sigma(\alpha)^i = \sum_i \sigma(c_i)\sigma(\alpha)^i = \sigma\left(\sum_i c_i\alpha^i\right).$$

Finally, we have $\sum_i c_i\alpha^i \in F(\alpha)$, and hence $x \in \sigma(F(\alpha))$, as proposed. ■

22.4 Normal extensions

Definition 22.6 Let $F \subset \mathbb{C}$ be a field. We call K/F a *normal extension* if K is closed under the process of taking conjugates over F, namely, whenever $\alpha \in K$, we have $\tilde{\alpha} \in K$ for all conjugates $\tilde{\alpha}$ of α over F.

Theorem 22.9 Let $F \subset \mathbb{C}$ be a field and let K/F be a finite extension. Then K is normal over F if and only if every embedding of K in $\mathbb{C}$ over F is in $\mathrm{Aut}(K/F)$.

Proof. We start with necessity. It is known that there is an element $\theta \in K$ such that $K = F(\theta)$. Let σ be an arbitrary embedding of K in $\mathbb{C}$ over F. By Lemma 22.1, $\sigma(\theta)$ is a conjugate of θ over F, and thus $\sigma(\theta) \in K$ since K/F is normal. It follows from Theorem 22.8 that $\sigma \in \mathrm{Aut}(K/F)$.

We then prove sufficiency. Let $\alpha \in K$ and assume that $\tilde{\alpha}$ is an arbitrary conjugate of α over F. Note that $F \subset F(\alpha) \subset K$. We know from Theorem 22.2 that there is an embedding τ of $F(\alpha)$ in $\mathbb{C}$ over F such that $\tau(\alpha) = \tilde{\alpha}$. We further extend τ to an embedding σ of K in $\mathbb{C}$. By our assumption, $\sigma \in \mathrm{Aut}(K/F)$. Noting that $\alpha \in F(\alpha)$, we conclude that $\tilde{\alpha} = \tau(\alpha) = \sigma(\alpha) \in K$, as required. ■

Theorem 22.10 Let $F \subset \mathbb{C}$ be a field. Let $\alpha, \beta, \ldots, \gamma$ be in $\mathbb{C}$ with $\alpha, \beta, \ldots, \gamma$ algebraic over F. If the conjugates of $\alpha, \beta, \ldots, \gamma$ over F are in $F(\alpha, \beta, \ldots, \gamma)$, then $F(\alpha, \beta, \ldots, \gamma)$ is normal over F.

Proof. We know from Corollary 21.14 that $F(\alpha, \beta, \ldots, \gamma) = F[\alpha, \beta, \ldots, \gamma]$. Let us write $K = F(\alpha, \beta, \ldots, \gamma)$. Recall that there is an element $\theta \in K$ such that $K = F(\theta)$. Meanwhile,

$\theta = f(\alpha, \beta, \ldots, \gamma)$ with $f \in F[x_\alpha, x_\beta, \ldots, x_\gamma]$ a multivariate polynomial over F. Letting σ be an arbitrary embedding of K in $\mathbb{C}$ over F, we have

$$\sigma(\theta) = \sigma\big(f(\alpha, \beta, \ldots, \gamma)\big) = f\big(\sigma(\alpha), \sigma(\beta), \ldots, \sigma(\gamma)\big).$$

By Lemma 22.1, $\sigma(\alpha)$ is a conjugate of α over F, so it is in K. The same argument also works for $\beta, \ldots, \gamma$. As a consequence, $\sigma(\theta) \in K = F(\theta)$. We conclude from Theorem 22.8 that $\sigma \in \mathrm{Aut}(K/F)$, and further from Theorem 22.9 that K/F is normal by recalling that σ is arbitrarily chosen. ■

Theorem 22.11 Let $F \subset \mathbb{C}$ be a field and let K/F be a finite extension. Then there is a finite extension L of K such that L is normal over F.

Proof. We still assume that the element $\theta \in K$ is such that $K = F(\theta)$. Now let $\theta_1, \ldots, \theta_n$ be all conjugates of θ over F. Then $L = F(\theta_1, \ldots, \theta_n)$ is the desired extension of K that is normal over F by Theorem 22.10. ■

■ **Example 22.2** Consider the extension $\mathbb{Q}(\sqrt[3]{3})/\mathbb{Q}$. Since $\sqrt[3]{3}$ has conjugates $\sqrt[3]{3}$, $\omega\sqrt[3]{3}$ and $\omega^2\sqrt[3]{3}$ over $\mathbb{Q}$ where $\omega = e^{2\pi i/3}$, we know that $\mathbb{Q}(\sqrt[3]{3})/\mathbb{Q}$ is not normal over $\mathbb{Q}$ for $\omega\sqrt[3]{3} \notin \mathbb{Q}(\sqrt[3]{3})$. However, we may then extend $\mathbb{Q}(\sqrt[3]{3})$ to $\mathbb{Q}(\sqrt[3]{3}, \omega\sqrt[3]{3}, \omega^2\sqrt[3]{3})$ to arrive at a normal extension of $\mathbb{Q}$. Observe that $\mathbb{Q}(\sqrt[3]{3}, \omega\sqrt[3]{3}, \omega^2\sqrt[3]{3}) = \mathbb{Q}(\sqrt[3]{3}, \omega\sqrt[3]{3})$ since $\omega^2\sqrt[3]{3} = \frac{(\omega\sqrt[3]{3})^2}{\sqrt[3]{3}}$. ■

22.5 Galois extensions

We have shown in Theorem 22.7 that for K/F a finite extension, $|\mathrm{Aut}(K/F)| \le [K:F]$. We are in particular interested in the case where $|\mathrm{Aut}(K/F)|$ reaches the largest possible value.

Definition 22.7 Let $F \subset \mathbb{C}$ be a field and let K/F be a finite extension. We call K/F a *Galois extension* if $|\mathrm{Aut}(K/F)| = [K:F]$. In this case, $\mathrm{Aut}(K/F)$ is called the *Galois group* of K/F, denoted by $\mathrm{Gal}(K/F)$.

Theorem 22.12 Let $F \subset \mathbb{C}$ be a field and let K/F be a finite extension. Then K is Galois over F if and only if K is normal over F.

Proof. Note that by Corollary 22.4, there are exactly $[K:F]$ distinct embeddings of K in $\mathbb{C}$ over F. Now it follows from Theorem 22.9 that K/F is normal if and only if $|\mathrm{Aut}(K/F)| = [K:F]$, i.e. K/F is Galois. ■

R The validity of Theorem 22.12 relies entirely on our assumption in Sect. 21.2 that every field is assumed to be a subfield of $\mathbb{C}$ under the usual addition and multiplication. This assumption ensures that the minimal polynomial of every algebraic α over F has **distinct** roots (Corollary 21.9), and hence that there are exactly $[K:F]$ **distinct** embeddings of K in $\mathbb{C}$ over F (Corollary 22.4).

The French mathematician Évariste Galois established a profound theory on the connection between field theory and group theory, known as the *Galois theory*, in which one of the most significant results is the *Fundamental Theorem of Galois Theory*. In its most basic form, the following statement is asserted, but we will not cover the proof in this series of notes.

Fundamental Theorem of Galois Theory There is a one-to-one correspondence between the intermediate fields of a Galois extension and the subgroups of its Galois group.

22.6 Comments on separability

In general, we may consider finite extensions K over a generic field F. We still say K/F is Galois if $|\mathrm{Aut}(K/F)| = [K : F]$. Now K/F is Galois if and only if

(i) K/F **is normal**: for every $\alpha \in K$, its minimal polynomial over F has all roots in K;

(ii) ***and*** K/F **is separable**: for every $\alpha \in K$, its minimal polynomial over F has all roots distinct in an algebraic closure of F.

Note that given a field F with additive identity 0 and multiplicative identity 1, we call the largest integer n such that

$$\underbrace{1+1+\cdots+1}_{n \text{ terms}} = 0$$

the *characteristic* of F. If it exists, then n is a prime (e.g. $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$); if it does not exist, we say F has characteristic 0 (e.g. $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$).

It is known that every algebraic extension of a field of characteristic zero is separable, and that every algebraic extension of a finite field is separable. The first claim explains why we are allowed to remove the separability condition in Theorem 22.12.

However, there *do* exist extensions that are *not* separable. One simple example is as follows. Let $\mathbb{F}_p(x)$ be the field of rational functions in the indeterminate x over $\mathbb{F}_p$. Consider the extension $K/F = \mathbb{F}_p(x)/\mathbb{F}_p(x^p)$. The element $x \in K$ has minimal polynomial $f(X) = X^p - x^p \in F[X]$ of degree p over F. However, over the extension field K, we also have $X^p - x^p = (X-x)^p$. Hence, as a polynomial in $K[X]$ of degree p, $f(X)$ has p repeated roots $X = x$. Thus, the extension $K/F = \mathbb{F}_p(x)/\mathbb{F}_p(x^p)$ is inseparable.

23. Algebraic integers

23.1 Integrality

In the previous lectures, rational numbers have been generalized to algebraic elements over a field. Now we shall focus on the analogy of integers. For example, if K is a number field, which elements in K should be viewed as "integral"?

Let us first introduce the concept of *module*, which can be viewed as a generalization of the notion of vector space in which the field of scalars is replaced by a ring.

Definition 23.1 Let R be a ring with 1 its multiplicative identity. A *(left) R-module M* consists of an abelian group $(M,+)$ and an operation $\circ : R \times M \to M$ such that for all $r,s \in R$ and $x,y \in M$,

(i) $r \circ (x+y) = r \circ x + r \circ y$;
(ii) $(r+s) \circ x = r \circ x + s \circ x$;
(iii) $(rs) \circ x = r \circ (s \circ x)$;
(iv) $1 \circ x = x$.

The operation $\circ$ is called *scalar multiplication*. Often the symbol $\circ$ is omitted.

Definition 23.2 An R-module M is *finitely generated* if there exist finitely many elements $x_1, \ldots, x_n \in M$ such that every $x \in M$ can be written as $x = \sum_{i=1}^{n} a_i x_i$ with $a_i \in R$.

▪ **Example 23.1** The set of 2-dimensional row vectors over $\mathbb{Z}$, $\{(x,y) : x,y \in \mathbb{Z}\}$, forms a $\mathbb{Z}$-module. Also, it is finitely generated by $\{(1,0),(0,1)\}$. ▪

Theorem 23.1 Let $A \subset B \subset C$ be three rings. If C is a finitely generated B-module and B is a finitely generated A-module, then C is a finitely generated A-module.

Proof. Since C is a finitely generated B-module, we can find elements $\gamma_1, \ldots, \gamma_m \in C$ such that every $c \in C$ can be written as $c = \sum_{j=1}^{m} b_j \gamma_j$ with $b_j \in B$. Also, since B is a finitely generated A-module, we can find elements $\beta_1, \ldots, \beta_n \in B$ such that each b_j can be written as $b_j = \sum_{i=1}^{n} a_{ij} \beta_i$ with $a_{ij} \in A$. Hence, $c = \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} (\beta_i \gamma_j)$. That is, C is generated by the finitely many elements $\{\beta_i \gamma_j \in C : 1 \le i \le n,\ 1 \le j \le m\}$ over A. ■

Definition 23.3 Let R be a ring. An element α is said to be *integral over R* if α is a root of a monic polynomial over R.

Theorem 23.2 (Criteria of Integrality). Let R be an integral domain and let A be a subring of R. Let $\alpha \in R$. Then the following statements are equivalent:

(i) α is integral over A;
(ii) The ring $A[\alpha]$ is a finitely generated A-module;
(iii) There is a finitely generated nonzero A-module $B \subset R$ such that $\alpha B \subset B$.

Proof. (i) ⇒ (ii): Assume that α is a root of $x^n + a_{n-1}x^{n-1} + \cdots + a_0 \in A[x]$. It is plain to see that $A[\alpha]$ is generated by $\{1, \alpha, \ldots, \alpha^{n-1}\}$ over A by the fact that we can write α^n as $\alpha^n = -a_{n-1}\alpha^{n-1} - \cdots - a_0$ where $a_i \in A$.

(ii) ⇒ (iii): Choose $B = A[\alpha]$.

(iii) ⇒ (i): Let $b_1, \ldots, b_m$ be elements in B such that $B = Ab_1 + \cdots + Ab_m$. Note that not all b_i are zero since B is nonzero. Now $\alpha B \subset B$ implies that $\alpha b_i = \sum_{j=1}^m a_{ij} b_j$ with $a_{ij} \in A$ for all i and j. In matrix form we have

$$\begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mm} \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} = \alpha \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}. \tag{23.1}$$

We may extend the integral domain R to its quotient field K (e.g. extending $\mathbb{Z}$ to $\mathbb{Q}$), and regard (23.1) with entries in K. The reason for doing so is that eigensystems are in general considered in the setting of vector spaces over a *field*. It follows that $(b_1, \ldots, b_m)^\mathsf{T}$ is an eigenvector of (a_{ij}) for eigenvalue α. Then α is a root of the characteristic polynomial of (a_{ij}), i.e. $\det\big(xI_m - (a_{ij})\big)$, which is monic over A since $a_{ij} \in A$. So α is integral over A. ■

Theorem 23.3 Let R be an integral domain and let A be a subring of R. Then the set of elements in R that are integral over A forms a ring.

Proof. It is plain that the additive inverse of any integral element over A is also integral. Now let $\alpha, \beta \in R$ be integral over A. It is sufficient to show that $\alpha + \beta$ and $\alpha\beta$ are integral over A. By assumption, $A[\alpha]$ is a finitely generated A-module. Also, β integral over A implies that β is integral over $A[\alpha] \subset R$, and hence that $A[\alpha, \beta] = A[\alpha][\beta]$ is a finitely generated $A[\alpha]$-module. It follows from Theorem 23.1 that $A[\alpha, \beta]$ is a finitely generated A-module. Finally, the finitely generated A-module $A[\alpha, \beta] \subset R$ is such that $(\alpha + \beta)A[\alpha, \beta] \subset A[\alpha, \beta]$ and $(\alpha\beta)A[\alpha, \beta] \subset A[\alpha, \beta]$. We conclude that $\alpha + \beta$ and $\alpha\beta$ are integral over A. ■

R For a constructive proof, the argument for Theorem 21.19 still works. For instance, in Example 21.2, $\alpha = \sqrt{2}$ and $\beta = \sqrt[3]{3}$ are indeed integral over $\mathbb{Z}$. Further, $s(\sqrt{2} + \sqrt[3]{3}) = 0$ and $p(\sqrt{2} \cdot \sqrt[3]{3}) = 0$ where $s(x) = x^6 - 6x^4 - 6x^3 + 12x^2 - 36x + 1$ and $p(x) = x^6 - 72$ are monic over $\mathbb{Z}$. It turns out that $\sqrt{2} + \sqrt[3]{3}$ and $\sqrt{2} \cdot \sqrt[3]{3}$ are also integral over $\mathbb{Z}$.

Definition 23.4 Let R be a ring. A set S is said to be *integral over* R if every element in S is integral over R.

Theorem 23.4 Let R be an integral domain and let A be a subring of R. If $\alpha \in R$ is integral over A, then the ring $A[\alpha]$ is integral over A.

Proof. This is an immediate consequence of Theorem 23.3. ■

We also have a parallel result to the transitivity of algebraicity of field extensions as described in Theorem 21.15.

Theorem 23.5 (Transitivity of Integrality). Let R be an integral domain. Let $A \subset B \subset C$ be three subrings of R. If C is integral over B and B is integral over A, then C is integral over A.

Proof. Let $c \in C$. Since c is integral over B, it is a root of $x^n + b_{n-1}x^{n-1} + \cdots + b_0$ with $b_i \in B$. Further, each b_i is integral over A, and as a consequence, $A[b_0, \ldots, b_{n-1}]$ is a finitely generated A-module. On the other hand, c is also integral over $A[b_0, \ldots, b_{n-1}]$. Hence, $A[b_0, \ldots, b_{n-1}, c] = A[b_0, \ldots, b_{n-1}][c]$ is a finitely generated $A[b_0, \ldots, b_{n-1}]$-module. By Theorem 23.1, $A[b_0, \ldots, b_{n-1}, c]$ is a finitely generated A-module. Finally, since $A[b_0, \ldots, b_{n-1}, c] \subset R$ is such that $cA[b_0, \ldots, b_{n-1}, c] \subset A[b_0, \ldots, b_{n-1}, c]$, we conclude that c is integral over A. ■

23.2 Algebraic integers

Definition 23.5 Let K be a number field. The set of all elements in K that are integral over $\mathbb{Z}$ forms a ring, denoted by $\mathscr{O}_K$ or $\mathbb{Z}_K$, called the *ring of (algebraic) integers* of K.

We check some basic properties of $\mathscr{O}_K$.

Proposition 23.6 We have $\mathscr{O}_{\mathbb{Q}} = \mathbb{Z}$.

Proof. It is clear that if $n \in \mathbb{Z}$, then n is a root of $x - n$, and hence n is integral over $\mathbb{Z}$. So $n \in \mathscr{O}_{\mathbb{Q}}$, thereby implying that $\mathbb{Z} \subset \mathscr{O}_{\mathbb{Q}}$. On the other hand, if $\alpha \in \mathscr{O}_{\mathbb{Q}}$, i.e. $\alpha \in \mathbb{Q}$ and α is a root of a monic polynomial over $\mathbb{Z}$, then by Theorem 19.5, $\alpha \in \mathbb{Z}$. That is, $\mathscr{O}_{\mathbb{Q}} \subset \mathbb{Z}$. Consequently, $\mathscr{O}_{\mathbb{Q}} = \mathbb{Z}$. ■

Proposition 23.7 Let K be a number field. Then $\mathscr{O}_K$ is integrally closed in K, i.e. if $\alpha \in K$ is integral over $\mathscr{O}_K$, then $\alpha \in \mathscr{O}_K$.

Proof. Note that $\mathbb{Z} \subset \mathscr{O}_K \subset \mathscr{O}_K[\alpha]$. Further, $\mathscr{O}_K$ is integral over $\mathbb{Z}$ by definition and $\mathscr{O}_K[\alpha]$ is integral over $\mathscr{O}_K$ by Theorem 23.4. Thus, Theorem 23.5 tells us that $\mathscr{O}_K[\alpha]$ is integral over $\mathbb{Z}$, and hence $\alpha \in \mathscr{O}_K[\alpha] \subset K$ is integral over $\mathbb{Z}$, i.e. $\alpha \in \mathscr{O}_K$. ■

Proposition 23.8 Let K be a number field and let L be a finite extension of K. Then $\mathscr{O}_L \supset \mathscr{O}_K$ and $\mathscr{O}_L \cap K = \mathscr{O}_K$. In particular, $\mathscr{O}_K \cap \mathbb{Q} = \mathscr{O}_{\mathbb{Q}} = \mathbb{Z}$.

Proof. We first have $\mathscr{O}_L \supset \mathscr{O}_K$ since $L \supset K$. Also, $\alpha \in \mathscr{O}_L \cap K$ if and only if α is integral over $\mathbb{Z}$, $\alpha \in L$ (for $\alpha \in \mathscr{O}_L$) and $\alpha \in K$. However, the simultaneous inclusion that $\alpha \in L$ and $\alpha \in K$ is equivalent to $\alpha \in L \cap K = K$. Hence, $\alpha \in \mathscr{O}_L \cap K$ if and only if $\alpha \in \mathscr{O}_K$. For the final relation, we use the extension $K/\mathbb{Q}$ for L/K and apply Proposition 23.6. ■

Theorem 23.9 Let α be an algebraic number. Then there is a nonzero integer $n \in \mathbb{Z}$ such that $n\alpha$ is integral over $\mathbb{Z}$.

Proof. Since α is algebraic over $\mathbb{Q}$, we assume that the minimal polynomial of α over $\mathbb{Q}$ is $p(x) = x^m + a_{m-1}x^{m-1} + \cdots + a_0$. Write each a_i in the irreducible expression $a_i = \frac{s_i}{r_i}$ with $r_i > 0$ and $(s_i, r_i) = 1$. Let $n = \operatorname{lcm}(r_0, \ldots, r_{m-1})$, so $r_i \mid n$ for every i. Define $f(x) = n^m p\left(\frac{x}{n}\right)$. Then

$$f(x) = n^m \left(\frac{x^m}{n^m} + a_{m-1}\frac{x^{m-1}}{n^{m-1}} + \cdots + a_0 \right) = x^m + \frac{s_{m-1}n}{r_{m-1}}x^{m-1} + \cdots + \frac{s_0 n^m}{r_0} \in \mathbb{Z}[x].$$

Further, $f(n\alpha) = n^m p(\alpha) = 0$. Hence, $n\alpha$ is integral over $\mathbb{Z}$. ■

Corollary 23.10 Let K be a number field. Then $K = \left\{\frac{\alpha}{n} : \alpha \in \mathscr{O}_K,\ n \in \mathbb{Z}\backslash\{0\}\right\}$.

Proof. We write $\frac{\mathscr{O}_K}{\mathbb{Z}} := \left\{\frac{\alpha}{n} : \alpha \in \mathscr{O}_K,\ n \in \mathbb{Z}\backslash\{0\}\right\}$. If $\frac{\alpha}{n} \in \frac{\mathscr{O}_K}{\mathbb{Z}}$, then $\frac{\alpha}{n} \in K$ since K is a field, while $\alpha \in \mathscr{O}_K \subset K$ and $n \in \mathbb{Z}\backslash\{0\} \subset K\backslash\{0\}$. If $\beta \in K$, then we may find a nonzero $n \in \mathbb{Z}$ such that $n\beta \in \mathscr{O}_K$. Now, $\beta = \frac{n\beta}{n} \in \frac{\mathscr{O}_K}{\mathbb{Z}}$. Hence, $K = \frac{\mathscr{O}_K}{\mathbb{Z}}$. ■

Theorem 23.11 Let K be a number field. If α is integral over $\mathbb{Z}$, so is every $\tilde{\alpha}$ conjugate to α over K. Also, the minimal polynomial of α over K is in $\mathscr{O}_K[x]$.

Proof. It is plain that α is algebraic over K. Let $p(x) = x^m + a_{m-1}x^{m-1} + \cdots + a_0 \in K[x]$ be the minimal polynomial of α over K. Then $p(\tilde{\alpha}) = 0$ as $\tilde{\alpha}$ is conjugate to α over K. Since α is integral over $\mathbb{Z}$, there is a polynomial $f(x) = x^n + b_{n-1}x^{n-1} + \cdots + b_0 \in \mathbb{Z}[x]$ such that $f(\alpha) = 0$. Note that $f(x)$ is also in $K[x]$. By Theorem 21.5, we can find a polynomial $q(x) \in K[x]$ such that $f(x) = p(x)q(x)$. Now, $f(\tilde{\alpha}) = p(\tilde{\alpha})q(\tilde{\alpha}) = 0$, thereby implying that $\tilde{\alpha}$ is integral over $\mathbb{Z}$.

Let $\alpha_1, \ldots, \alpha_m$ be the roots of $p(x)$ in $\mathbb{C}$, that is, $\alpha_1, \ldots, \alpha_m$ are the conjugates of α over K, so that they are integral over $\mathbb{Z}$. Note that $x^m + a_{m-1}x^{m-1} + \cdots + a_0 = p(x) = (x-\alpha_1)\cdots(x-\alpha_m)$. Hence, each coefficient a_i can be written as $a_i = g_i(\alpha_1, \ldots, \alpha_m)$ where g_i is a certain multivariate polynomial in $\mathbb{Z}[x_1, \ldots, x_m]$. It follows from Theorem 23.3 that a_i is also integral over $\mathbb{Z}$ for every i. However, $p(x) \in K[x]$ means that $a_i \in K$. Finally, by definition we have $a_i \in \mathscr{O}_K$, and thus $p(x) \in \mathscr{O}_K[x]$. ■

Corollary 23.12 Let α be integral over $\mathbb{Z}$. Then so is any $\tilde{\alpha}$ conjugate to α over $\mathbb{Q}$, and the minimal polynomial of α over $\mathbb{Q}$ is in $\mathbb{Z}[x]$.

Proof. We take $K = \mathbb{Q}$ in Theorem 23.11. ■

23.3 Trace and norm

Definition 23.6 Let K be a number field and let L be a finite extension of K of degree n. Let $\sigma_1, \ldots, \sigma_n$ be the n embeddings of L in $\mathbb{C}$ over K. For $\alpha \in L$, we define the *trace* and *norm* of α over K as

$$\mathrm{Tr}_{L/K}(\alpha) := \sum_{i=1}^{n} \sigma_i(\alpha), \tag{23.2}$$

$$N_{L/K}(\alpha) := \prod_{i=1}^{n} \sigma_i(\alpha). \tag{23.3}$$

The following properties are immediate by definition.

Proposition 23.13 Let K be a number field and let L be a finite extension of K of degree n. For any $\alpha, \beta \in L$ and $\delta \in K$, we have

(i) $\mathrm{Tr}_{L/K}(\alpha+\beta) = \mathrm{Tr}_{L/K}(\alpha) + \mathrm{Tr}_{L/K}(\beta)$;
(ii) $N_{L/K}(\alpha\beta) = N_{L/K}(\alpha)N_{L/K}(\beta)$;
(iii) $\mathrm{Tr}_{L/K}(\delta\alpha) = \delta\,\mathrm{Tr}_{L/K}(\alpha)$;
(iv) $N_{L/K}(\delta\alpha) = \delta^n\,\mathrm{Tr}_{L/K}(\alpha)$.

Proof. Let σ be an arbitrary embedding of L in $\mathbb{C}$ over K. For (i) and (ii), we use the facts that $\sigma(\alpha+\beta) = \sigma(\alpha) + \sigma(\beta)$ and $\sigma(\alpha\beta) = \sigma(\alpha)\sigma(\beta)$. For (iii) and (iv), we use the fact that $\sigma(\delta\alpha) = \sigma(\delta)\sigma(\alpha) = \delta\sigma(\alpha)$ since σ fixes K. ■

Now we consider further expressions of the trace and norm.

Lemma 23.14 Let K be a number field. Let α be algebraic over K with minimal polynomial $p(x) = x^r + a_{r-1}x^{r-1} + \cdots + a_0$. Then

$$\mathrm{Tr}_{K(\alpha)/K}(\alpha) = -a_{r-1} \tag{23.4}$$

and

$$N_{K(\alpha)/K}(\alpha) = (-1)^r a_0. \tag{23.5}$$

In particular, $\mathrm{Tr}_{K(\alpha)/K}(\alpha)$ and $N_{K(\alpha)/K}(\alpha)$ are in K.

Proof. By Theorem 22.2, the $[K(\alpha):K] = r$ embeddings $\tau_1, \dots, \tau_r$ of $K(\alpha)$ in $\mathbb{C}$ over K are those such that $\tau_i(\alpha) = \alpha_i$ where $\alpha_1, \dots, \alpha_r$ are the conjugates of α over K. Note that

$$\begin{aligned} p(x) &= x^r + a_{r-1}x^{r-1} + \cdots + a_0 & \qquad (a_i \in K) \\ &= (x-\alpha_1)\cdots(x-\alpha_r) & \qquad (\alpha_i \in \mathbb{C}). \end{aligned}$$

Hence,

$$\mathrm{Tr}_{K(\alpha)/K}(\alpha) = \sum_{i=1}^{r} \tau_i(\alpha) = \sum_{i=1}^{r} \alpha_i = -a_{r-1}$$

and

$$N_{K(\alpha)/K}(\alpha) = \prod_{i=1}^{r} \tau_i(\alpha) = \prod_{i=1}^{r} \alpha_i = (-1)^r a_0,$$

as desired. ■

Theorem 23.15 Let K be a number field and let L be a finite extension of K of degree n. Let $\alpha \in L$ with minimal polynomial $p(x) = x^r + a_{r-1}x^{r-1} + \cdots + a_0$ over K. Then

$$\mathrm{Tr}_{L/K}(\alpha) = [L:K(\alpha)]\,\mathrm{Tr}_{K(\alpha)/K}(\alpha) = -\frac{na_{r-1}}{r} \tag{23.6}$$

and

$$N_{L/K}(\alpha) = N_{K(\alpha)/K}(\alpha)^{[L:K(\alpha)]} = (-1)^n a_0^{n/r}. \tag{23.7}$$

Consequently, $\mathrm{Tr}_{L/K}(\alpha)$ and $N_{L/K}(\alpha)$ are in K.

Furthermore, if $\alpha \in \mathscr{O}_L$, then $\mathrm{Tr}_{L/K}(\alpha)$ and $N_{L/K}(\alpha)$ are in $\mathscr{O}_K$. In this case, we further have that $\mathrm{Tr}_{L/\mathbb{Q}}(\alpha)$ and $N_{L/\mathbb{Q}}(\alpha)$ are in $\mathbb{Z}$.

Finally, $N_{L/K}(\alpha) = 0$ if and only if $\alpha = 0$.

Proof. Note that $K \subset K(\alpha) \subset L$. With the same notation as in the proof of Lemma 23.14, we extend each embedding τ_i to $[L:K(\alpha)] = \frac{n}{r}$ embeddings $\tau_{i,1}, \dots, \tau_{i,\frac{n}{r}}$ of L in $\mathbb{C}$ by Theorem 22.3. Now, for every i and j, $\tau_{i,j}(\alpha) = \tau_i(\alpha)$ since $\alpha \in K(\alpha)$. Hence,

$$\mathrm{Tr}_{L/K}(\alpha) = \sum_{i=1}^{r}\sum_{j=1}^{n/r} \tau_{i,j}(\alpha) = \sum_{i=1}^{r}\sum_{j=1}^{n/r} \tau_i(\alpha) = \frac{n}{r}\sum_{i=1}^{r} \alpha_i = -\frac{na_{r-1}}{r}$$

and

$$N_{L/K}(\alpha) = \prod_{i=1}^{r}\prod_{j=1}^{n/r} \tau_{i,j}(\alpha) = \prod_{i=1}^{r}\prod_{j=1}^{n/r} \tau_i(\alpha) = \left(\prod_{i=1}^{r} \alpha_i\right)^{n/r} = ((-1)^r a_0)^{n/r} = (-1)^n a_0^{n/r}.$$

We further note that $\alpha \in \mathscr{O}_L$ means that α is integral over $\mathbb{Z}$. By Theorem 23.11, the minimal polynomial $p(x)$ of α over K is in $\mathscr{O}_K[x]$, and hence a_0 and a_{r-1} are in $\mathscr{O}_K$. It follows that $\mathrm{Tr}_{L/K}(\alpha)$ and $N_{L/K}(\alpha)$ are in $\mathscr{O}_K$ as $\frac{n}{r} = [L : K(\alpha)]$ is an integer.

Finally, $N_{L/K}(\alpha) = 0$ if and only if $a_0 = 0$. However, in this case, we observe that $p(x) = x^r + a_{r-1}x^{r-1} + \cdots + a_1 x$ is divisible by x over K. Since $p(x)$ is irreducible over K, the only possibility is $p(x) = x$, which is equivalent to $\alpha = 0$. ■

Theorem 23.16 Let K, L and M be number fields with $K \subset L \subset M$. Then for all $\alpha \in M$,

$$\mathrm{Tr}_{M/K}(\alpha) = \mathrm{Tr}_{L/K}\left(\mathrm{Tr}_{M/L}(\alpha)\right) \tag{23.8}$$

and

$$N_{M/K}(\alpha) = N_{L/K}\left(N_{M/L}(\alpha)\right). \tag{23.9}$$

Proof. Suppose that $[M : L] = m$ with embeddings of M in $\mathbb{C}$ over L given by $\sigma_1, \ldots, \sigma_m$, and that $[L : K] = n$ with embeddings of L in $\mathbb{C}$ over K given by $\tau_1, \ldots, \tau_n$. Given $\alpha \in M$, we have

$$\mathrm{Tr}_{L/K}\left(\mathrm{Tr}_{M/L}(\alpha)\right) = \mathrm{Tr}_{L/K}\left(\sum_{i=1}^{m} \sigma_i(\alpha)\right) = \sum_{j=1}^{n} \tau_j\left(\sum_{i=1}^{m} \sigma_i(\alpha)\right)$$

and

$$N_{L/K}\left(N_{M/L}(\alpha)\right) = N_{L/K}\left(\prod_{i=1}^{m} \sigma_i(\alpha)\right) = \prod_{j=1}^{n} \tau_j\left(\prod_{i=1}^{m} \sigma_i(\alpha)\right).$$

We would like to compose τ_j and σ_i, but we cannot do so directly since the image of σ_i may not lie in the domain of τ_j. To overcome this issue, we need to find a finite Galois extension G/K such that $M \subset G$. This is doable by Theorem 22.11. Now, all embeddings of G in $\mathbb{C}$ over K are given by elements in $\mathsf{Gal}(G/K)$. Let us extend σ_i to an embedding $\tilde{\sigma}_i$ of G in $\mathbb{C}$ over L (and hence over K) for each i and extend τ_j to an embedding $\tilde{\tau}_j$ of G in $\mathbb{C}$ over K for each j. Then $\tilde{\sigma}_i, \tilde{\tau}_j \in \mathsf{Gal}(G/K)$, and thus we can compose $\tilde{\tau}_j$ with $\tilde{\sigma}_i$, thereby getting a new embedding of G in $\mathbb{C}$ over K, namely, $\tilde{\tau}_j \circ \tilde{\sigma}_i$.

By Corollary 22.4, there are $[M : K] = mn$ embeddings of M in $\mathbb{C}$ over K. We claim that they are given by the mn restricted embeddings $\tilde{\tau}_j \circ \tilde{\sigma}_i\big|_M$ for $1 \le i \le m$ and $1 \le j \le n$. For this, it is sufficient to show that these restricted embeddings are distinct. Supposing $\tilde{\tau}_j \circ \tilde{\sigma}_i\big|_M = \tilde{\tau}_{j'} \circ \tilde{\sigma}_{i'}\big|_M$, we want to show that $i = i'$ and $j = j'$. Let $x \in L$ be arbitrary. Recalling that $\tilde{\sigma}_i$ and $\tilde{\sigma}_{i'}$ fix L, we have

$$\tau_j(x) = \tilde{\tau}_j(x) = \tilde{\tau}_j \circ \tilde{\sigma}_i(x) = \tilde{\tau}_j \circ \tilde{\sigma}_i\big|_M(x) = \tilde{\tau}_{j'} \circ \tilde{\sigma}_{i'}\big|_M(x) = \tilde{\tau}_{j'} \circ \tilde{\sigma}_{i'}(x) = \tilde{\tau}_{j'}(x) = \tau_{j'}(x).$$

Hence, $j = j'$. Now let $y \in M$ be arbitrary. Since

$$\tilde{\tau}_j \circ \tilde{\sigma}_i(y) = \tilde{\tau}_j \circ \tilde{\sigma}_i\big|_M(y) = \tilde{\tau}_{j'} \circ \tilde{\sigma}_{i'}\big|_M(y) = \tilde{\tau}_j \circ \tilde{\sigma}_{i'}\big|_M(y) = \tilde{\tau}_j \circ \tilde{\sigma}_{i'}(y),$$

and $\tilde{\tau}_j \in \mathsf{Gal}(G/K)$ is one-to-one, we have $\tilde{\sigma}_i(y) = \tilde{\sigma}_{i'}(y)$ and hence $i = i'$.

In conclusion,

$$\mathrm{Tr}_{L/K}\left(\mathrm{Tr}_{M/L}(\alpha)\right) = \sum_{j=1}^{n} \tilde{\tau}_j\left(\sum_{i=1}^{m} \tilde{\sigma}_i(\alpha)\right) = \sum_{j=1}^{n}\sum_{i=1}^{m} \tilde{\tau}_j \circ \tilde{\sigma}_i(\alpha) = \sum_{j=1}^{n}\sum_{i=1}^{m} \tilde{\tau}_j \circ \tilde{\sigma}_i\big|_M(\alpha) = \mathrm{Tr}_{M/K}(\alpha)$$

and

$$N_{L/K}\big(N_{M/L}(\alpha)\big) = \prod_{j=1}^{n} \tilde{\tau}_j \left(\prod_{i=1}^{m} \tilde{\sigma}_i(\alpha) \right) = \prod_{j=1}^{n}\prod_{i=1}^{m} \tilde{\tau}_j \circ \tilde{\sigma}_i(\alpha) = \prod_{j=1}^{n}\prod_{i=1}^{m} \tilde{\tau}_j \circ \tilde{\sigma}_i\big|_M(\alpha) = N_{M/K}(\alpha),$$

as proposed. ■

24. Discriminant

24.1 Discriminant

Definition 24.1 Let $K \subset \mathbb{C}$ be a field and let L be a finite extension of K of degree n. Let $\sigma_1, \ldots, \sigma_n$ be the n embeddings of L in $\mathbb{C}$ over K. Given $\alpha_1, \ldots, \alpha_n \in L$, we define the *discriminant* of the n-tuple $(\alpha_1, \ldots, \alpha_n)$ by

$$\operatorname{disc}(\alpha_1, \ldots, \alpha_n) := \left(\det \begin{pmatrix} \sigma_1(\alpha_1) & \cdots & \sigma_1(\alpha_n) \\ \vdots & \ddots & \vdots \\ \sigma_n(\alpha_1) & \cdots & \sigma_n(\alpha_n) \end{pmatrix}\right)^2 .$$

R Observe that $\operatorname{disc}(\alpha_1, \ldots, \alpha_n)$ is independent of the order of $\alpha_1, \ldots, \alpha_n$, as well as the order of the embeddings $\sigma_1, \ldots, \sigma_n$.

Let α be algebraic over $K \subset \mathbb{C}$, of degree n. We know from Theorem 21.11 that $\{1, \alpha, \ldots, \alpha^{n-1}\}$ forms a basis for $K(\alpha)$ over K.

Definition 24.2 Let $K \subset \mathbb{C}$ be a field. Let α be algebraic over K of degree n. Let $\sigma_1, \ldots, \sigma_n$ be the n embeddings of $K(\alpha)$ in $\mathbb{C}$ over K. We define the *discriminant* of α by

$$\operatorname{disc}(\alpha) = \operatorname{disc}(1, \alpha, \ldots, \alpha^{n-1}).$$

Theorem 24.1 Let $K \subset \mathbb{C}$ be a field. Let α be algebraic over K of degree n with minimal polynomial $p(x) \in K[x]$. Then

$$\operatorname{disc}(\alpha) = (-1)^{\frac{n(n-1)}{2}} N_{K(\alpha)/K}\big(p'(\alpha)\big), \tag{24.1}$$

where $p'(x)$ is the derivative of $p(x)$.

Proof. Assume that $\alpha_1, \ldots, \alpha_n$ are the conjugates of α over K. Then the n embeddings $\sigma_1, \ldots, \sigma_n$ of $K(\alpha)$ in $\mathbb{C}$ over K send α to $\alpha_1, \ldots, \alpha_n$, respectively. Now,

$$\operatorname{disc}(\alpha) = \Big(\det\big(\sigma_i(\alpha^k)\big)\Big)^2 = \Big(\det\big(\alpha_i^k\big)\Big)^2 .$$

Noting that the $n \times n$ square matrix

$$(\alpha_i^k) = \begin{pmatrix} 1 & \alpha_1 & \cdots & \alpha_1^{n-1} \\ 1 & \alpha_2 & \cdots & \alpha_2^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \alpha_n & \cdots & \alpha_n^{n-1} \end{pmatrix}$$

is a Vandermonde matrix, we have

$$\operatorname{disc}(\alpha) = \left(\prod_{1 \le i < j \le n} (\alpha_j - \alpha_i) \right)^2 = (-1)^{\frac{n(n-1)}{2}} \prod_{\substack{1 \le i,j \le n \\ i \ne j}} (\alpha_i - \alpha_j).$$

Recall that $p(x) = (x - \alpha_1) \cdots (x - \alpha_n)$. We have

$$p'(\alpha_i) = \lim_{x \to \alpha_i} \frac{p(x)}{x - \alpha_i} = \prod_{\substack{1 \le j \le n \\ j \ne i}} (\alpha_i - \alpha_j).$$

Thus,

$$\operatorname{disc}(\alpha) = (-1)^{\frac{n(n-1)}{2}} \prod_{1 \le i \le n} p'(\alpha_i) = (-1)^{\frac{n(n-1)}{2}} \prod_{1 \le i \le n} \sigma_i\big(p'(\alpha)\big) = (-1)^{\frac{n(n-1)}{2}} N_{K(\alpha)/K}\big(p'(\alpha)\big),$$

as desired. ■

■ **Example 24.1** Let $K = \mathbb{Q}$ and $\alpha = \sqrt[3]{2}$. The minimal polynomial of $\sqrt[3]{2}$ over $\mathbb{Q}$ is $p(x) = x^3 - 2$. Then $p'(x) = 3x^2$. So

$$\begin{aligned} \operatorname{disc}(\sqrt[3]{2}) &= (-1)^{\frac{3(3-1)}{2}} N_{\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q}}\big(p'(\sqrt[3]{2})\big) = -N_{\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q}}\big(3 \cdot (\sqrt[3]{2})^2\big) \\ &= -27 N_{\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q}}\big(\sqrt[3]{2}\big)^2 = -27\big(\sqrt[3]{2} \cdot \sqrt[3]{2} e^{\frac{2\pi i}{3}} \cdot \sqrt[3]{2} e^{\frac{4\pi i}{3}}\big)^2 = -27 \cdot 4. \end{aligned}$$

Hence, $\operatorname{disc}(\sqrt[3]{2}) = -108$. ■

Theorem 24.2 Let $K \subset \mathbb{C}$ be a field and let L be a finite extension of K of degree n. Let $\alpha_1, \ldots, \alpha_n \in L$. Then

$$\operatorname{disc}(\alpha_1, \ldots, \alpha_n) = \det \begin{pmatrix} \operatorname{Tr}_{L/K}(\alpha_1 \alpha_1) & \cdots & \operatorname{Tr}_{L/K}(\alpha_1 \alpha_n) \\ \vdots & \ddots & \vdots \\ \operatorname{Tr}_{L/K}(\alpha_n \alpha_1) & \cdots & \operatorname{Tr}_{L/K}(\alpha_n \alpha_n) \end{pmatrix}. \tag{24.2}$$

Consequently, $\operatorname{disc}(\alpha_1, \ldots, \alpha_n) \in K$. Further, if $\alpha_1, \ldots, \alpha_n \in \mathscr{O}_L$, then $\operatorname{disc}(\alpha_1, \ldots, \alpha_n) \in \mathscr{O}_K$.

Proof. We shall use the fact that for any square matrix A, $\det A^{\mathsf{T}} = \det A$. Let $\sigma_1, \ldots, \sigma_n$ be the n embeddings of L in $\mathbb{C}$ over K. Then

$$\begin{aligned} \operatorname{disc}(\alpha_1, \ldots, \alpha_n) &= \det\big(\sigma_i(\alpha_j)\big)^2 = \det\big(\sigma_i(\alpha_j)\big)^{\mathsf{T}} \det\big(\sigma_i(\alpha_j)\big) \\ &= \det\big(\sigma_i(\alpha_j)\big)^{\mathsf{T}}\big(\sigma_i(\alpha_j)\big) = \det\big(\sigma_j(\alpha_i)\big)\big(\sigma_i(\alpha_j)\big). \end{aligned}$$

Note that the (i,j)-th entry of the matrix product $\big(\sigma_j(\alpha_i)\big)\big(\sigma_i(\alpha_j)\big)$ is

$$\sum_{h=1}^{n} \sigma_h(\alpha_i)\sigma_h(\alpha_j) = \sum_{h=1}^{n} \sigma_h(\alpha_i\alpha_j) = \operatorname{Tr}_{L/K}(\alpha_i\alpha_j).$$

The desired result then follows. ■

Theorem 24.3 Let $K \subset \mathbb{C}$ be a field and let L be a finite extension of K of degree n. Let $\alpha_1, \ldots, \alpha_n \in L$. If $\beta_1, \ldots, \beta_n \in L$ are such that

$$(\beta_1, \ldots, \beta_n) = (\alpha_1, \ldots, \alpha_n) \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nn} \end{pmatrix}, \tag{24.3}$$

where $\big(c_{ij}\big)_{1 \le i,j \le n} \in \mathrm{Mat}_{n,n}(K)$, then

$$\operatorname{disc}(\beta_1, \ldots, \beta_n) = \big(\det\big(c_{ij}\big)\big)^2 \cdot \operatorname{disc}(\alpha_1, \ldots, \alpha_n). \tag{24.4}$$

Proof. Let $\sigma_1, \ldots, \sigma_n$ be the n embeddings of L in $\mathbb{C}$ over K. For each i, applying σ_i to (24.3) and recalling that σ_i fixes K, we have

$$\big(\sigma_i(\beta_1), \ldots, \sigma_i(\beta_n)\big) = \big(\sigma_i(\alpha_1), \ldots, \sigma_i(\alpha_n)\big) \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nn} \end{pmatrix}.$$

Hence,

$$\begin{pmatrix} \sigma_1(\beta_1) & \cdots & \sigma_1(\beta_n) \\ \vdots & \ddots & \vdots \\ \sigma_n(\beta_1) & \cdots & \sigma_n(\beta_n) \end{pmatrix} = \begin{pmatrix} \sigma_1(\alpha_1) & \cdots & \sigma_1(\alpha_n) \\ \vdots & \ddots & \vdots \\ \sigma_n(\alpha_1) & \cdots & \sigma_n(\alpha_n) \end{pmatrix} \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nn} \end{pmatrix}.$$

The desired relation follows from the definition of the discriminant. ■

24.2 Linear independence of elements in a field extension

Discriminant serves as a useful tool to characterize the linear independence of elements in a field extension.

Theorem 24.4 Let $K \subset \mathbb{C}$ be a field and let L be a finite extension of K of degree n. Let $\alpha_1, \ldots, \alpha_n \in L$. Then $\alpha_1, \ldots, \alpha_n$ are linearly dependent over K if and only if $\operatorname{disc}(\alpha_1, \ldots, \alpha_n) = 0$.

Proof. Let $\sigma_1, \ldots, \sigma_n$ be the n embeddings of L in $\mathbb{C}$ over K.

We start with necessity. Suppose that $\alpha_1, \ldots, \alpha_n$ are linearly dependent over K. Then there exist $a_1, \ldots, a_n \in K$, not all zero, such that

$$a_1\alpha_1 + \cdots + a_n\alpha_n = 0.$$

Applying σ_i for each i to the above and noting that σ_i fixes K, we have

$$a_1\sigma_i(\alpha_1) + \cdots + a_n\sigma_i(\alpha_n) = 0.$$

In matrix form, we obtain

$$\begin{pmatrix} \sigma_1(\alpha_1) & \cdots & \sigma_1(\alpha_n) \\ \vdots & \ddots & \vdots \\ \sigma_n(\alpha_1) & \cdots & \sigma_n(\alpha_n) \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Recall that $(a_1, \ldots, a_n)^{\mathsf{T}}$ is a nonzero column vector. Hence, $\det\big(\sigma_i(\alpha_j)\big) = 0$, namely, $\operatorname{disc}(\alpha_1, \ldots, \alpha_n) = 0$.

We then prove sufficiency. It is known from Theorem 24.2 that $\mathsf{disc}(\alpha_1,\dots,\alpha_n) = \det\big(\mathrm{Tr}_{L/K}(\alpha_i\alpha_j)\big)$. Supposing that $\alpha_1,\dots,\alpha_n$ are linearly independent over K, we shall show that $\mathsf{disc}(\alpha_1,\dots,\alpha_n) \neq 0$. Our starting point is the map

$$\begin{array}{cccc} (\cdot,\cdot): & L\times L & \to & K \\ & (\alpha,\beta) & \mapsto & \mathrm{Tr}_{L/K}(\alpha\beta) \end{array}$$

Note that $(\cdot,\cdot)$ is K-bilinear (i.e. $(a_1\alpha_1+a_2\alpha_2, b_1\beta_1+b_2\beta_2) = a_1b_1(\alpha_1,\beta_1)+a_1b_2(\alpha_1,\beta_2)+a_2b_1(\alpha_2,\beta_1)+a_2b_2(\alpha_2,\beta_2)$ for all $a_1,a_2,b_1,b_2\in K$ and $\alpha_1,\alpha_2,\beta_1,\beta_2\in L$) and symmetric (i.e. $(\alpha,\beta)=(\beta,\alpha)$ for all $\alpha,\beta\in L$). We call this map a *pairing*. Observe that this pairing in *nondegenerate*, that is, for any $\beta\in L\backslash\{0\}$, there exists $\alpha\in L$ such that $(\alpha,\beta)\neq 0$. For instance, we can take $\alpha=\beta^{-1}$ and get $(\alpha,\beta)=(\beta^{-1},\beta)=\mathrm{Tr}_{L/K}(\beta^{-1}\beta)=\mathrm{Tr}_{L/K}(1)=n\neq 0$ by Theorem 23.15. Since we have assumed that $\alpha_1,\dots,\alpha_n$ are linearly independent over K, then given any $\alpha,\beta\in L$, we can uniquely write

$$\begin{cases} \alpha = a_1\alpha_1+\cdots+a_n\alpha_n, \\ \beta = b_1\alpha_1+\cdots+b_n\alpha_n. \end{cases}$$

Define the matrix $A=\big(\mathrm{Tr}_{L/K}(\alpha_i\alpha_j)\big)_{1\le i,j\le n}$. Then

$$(\alpha,\beta) = (a_1,\dots,a_n)\,A\begin{pmatrix} b_1\\ \vdots \\ b_n\end{pmatrix}.$$

If $\mathsf{disc}(\alpha_1,\dots,\alpha_n)=\det A=0$ where we make use of Theorem 24.2, then there exists a column vector $\begin{pmatrix} b_1\\ \vdots \\ b_n\end{pmatrix}\neq\begin{pmatrix} 0\\ \vdots \\ 0\end{pmatrix}$ such that $A\begin{pmatrix} b_1\\ \vdots \\ b_n\end{pmatrix}=\begin{pmatrix} 0\\ \vdots \\ 0\end{pmatrix}$. Now for $\beta=b_1\alpha_1+\cdots+b_n\alpha_n\neq 0$, we have $(\alpha,\beta)=0$ for all $\alpha\in L$. But this contradicts the nondegenerancy of $(\cdot,\cdot)$. ■

24.3 Integral bases

Let K be a number field of degree n over $\mathbb{Q}$. Then K as a vector space over $\mathbb{Q}$ can be spanned by a basis $\{\alpha_1,\dots,\alpha_n\}$ where $\alpha_1,\dots,\alpha_n\in K$ are linearly independent.

Lemma 24.5 Let K be a number field. There is a basis for K over $\mathbb{Q}$ such that its elements are in $\mathscr{O}_K$.

Proof. Let $\theta\in K$ be such that $K=\mathbb{Q}(\theta)$. Note that θ is algebraic over $\mathbb{Q}$. By Theorem 23.9, there is a certain nonzero $m\in\mathbb{Z}$ such that $\zeta:=m\theta\in\mathscr{O}_K$. Then $\mathbb{Q}(\zeta)=\mathbb{Q}(m\theta)=\mathbb{Q}(\theta)=K$. Assuming that $n=[K:\mathbb{Q}]$, then $\{1,\zeta,\dots,\zeta^{n-1}\}$ is a desired basis. ■

Theorem 24.6 Let K be a number field of degree n over $\mathbb{Q}$. There is a basis $\{\omega_1,\dots,\omega_n\}$ for K over $\mathbb{Q}$ with $\omega_i\in\mathscr{O}_K$ for all i such that every $\gamma\in\mathscr{O}_K$ has a unique representation $\gamma=m_1\omega_1+\cdots+m_n\omega_n$ with $m_i\in\mathbb{Z}$ for all i.

Proof. For any basis $\{\alpha_1,\dots,\alpha_n\}$ for K over $\mathbb{Q}$ with elements in $\mathscr{O}_K$, which exists by Lemma 24.5, we know from Theorem 24.2 that $\mathsf{disc}(\alpha_1,\dots,\alpha_n)\in\mathscr{O}_{\mathbb{Q}}=\mathbb{Z}$. Also, since $\alpha_1,\dots,\alpha_n$ are linearly independent, we have $\mathsf{disc}(\alpha_1,\dots,\alpha_n)\neq 0$ by Theorem 24.4. Assume that $\{\omega_1,\dots,\omega_n\}$ is such a basis with $|\mathsf{disc}(\omega_1,\dots,\omega_n)|$ minimal. We claim that this basis is as desired.

For any $\gamma \in \mathscr{O}_K$, since $\{\omega_1, \dots, \omega_n\}$ is a basis for K over $\mathbb{Q}$, we may uniquely write $\gamma = m_1\omega_1 + \cdots + m_n\omega_n$ with $m_i \in \mathbb{Q}$ for all i. We shall show that these m_i are indeed in $\mathbb{Z}$. If not, we assume, without loss of generality, that $m_1 \notin \mathbb{Z}$. Then we write $m_1 = m + r$ with $0 < r < 1$; here $m = \lfloor m_1 \rfloor$, the largest integer not exceeding m_1. Now we put

$$\omega_1' := \gamma - m\omega_1 = (m_1 - m)\omega_1 + m_2\omega_2 + \cdots + m_n\omega_n$$

and for $2 \le i \le n$,

$$\omega_i' := \omega_i.$$

Note that $\{\omega_1', \dots, \omega_n'\}$ also forms a basis for K over $\mathbb{Q}$ with elements in $\mathscr{O}_K$. Further,

$$\big(\omega_1', \omega_2', \dots, \omega_n'\big) = \big(\omega_1, \omega_2, \dots, \omega_n\big) \begin{pmatrix} m_1 - m & 0 & \cdots & 0 \\ m_2 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ m_n & 0 & \cdots & 1 \end{pmatrix}.$$

Noting that the determinant of the above square matrix is $m_1 - m = r$, we deduce from Theorem 24.3 that

$$\operatorname{disc}(\omega_1', \dots, \omega_n') = r^2 \operatorname{disc}(\omega_1, \dots, \omega_n).$$

However, since $0 < r < 1$, we have $|\operatorname{disc}(\omega_1', \dots, \omega_n')| = r^2 |\operatorname{disc}(\omega_1, \dots, \omega_n)| < |\operatorname{disc}(\omega_1, \dots, \omega_n)|$, violating the minimality of $|\operatorname{disc}(\omega_1, \dots, \omega_n)|$. Thus, we must have $m_i \in \mathbb{Z}$ for all i. ■

Definition 24.3 Let K be a number field of degree n over $\mathbb{Q}$. If a basis $\{\omega_1, \dots, \omega_n\}$ for K over $\mathbb{Q}$ is such that $\omega_i \in \mathscr{O}_K$ for all i, and that every $\gamma \in \mathscr{O}_K$ has a unique representation $\gamma = m_1\omega_1 + \cdots + m_n\omega_n$ with $m_i \in \mathbb{Z}$ for all i, then $\{\omega_1, \dots, \omega_n\}$ is called an *integral basis for* $\mathscr{O}_K$.

Theorem 24.7 Let K be a number field. All integral bases for $\mathscr{O}_K$ have the same discriminant.

Proof. Let $\{\alpha_1, \dots, \alpha_n\}$ and $\{\beta_1, \dots, \beta_n\}$ be two integral bases for $\mathscr{O}_K$. By definition, there are two matrices $A, B \in \mathrm{Mat}_{n,n}(\mathbb{Z})$ such that

$$\begin{aligned} \big(\beta_1, \dots, \beta_n\big) &= \big(\alpha_1, \dots, \alpha_n\big) \cdot A, \\ \big(\alpha_1, \dots, \alpha_n\big) &= \big(\beta_1, \dots, \beta_n\big) \cdot B. \end{aligned}$$

Hence,

$$\big(\alpha_1, \dots, \alpha_n\big) = \big(\alpha_1, \dots, \alpha_n\big) \cdot AB,$$

yielding that $AB = I_n$. It follows that $\det A \cdot \det B = 1$. Since $A, B \in \mathrm{Mat}_{n,n}(\mathbb{Z})$ and hence $\det A, \det B \in \mathbb{Z}$, we find that $\det A = \det B = \pm 1$. Finally, by Theorem 24.3, we conclude that

$$\operatorname{disc}(\beta_1, \dots, \beta_n) = (\det A)^2 \cdot \operatorname{disc}(\alpha_1, \dots, \alpha_n) = (\pm 1)^2 \operatorname{disc}(\alpha_1, \dots, \alpha_n) = \operatorname{disc}(\alpha_1, \dots, \alpha_n),$$

as required. ■

Integral bases allow us to define an invariant for a number field.

Definition 24.4 Let K be a number field. We define the *discriminant* of K, denoted by d_K, as the discriminant of an integral basis for $\mathscr{O}_K$. In particular, $d_K \in \mathbb{Z} \backslash \{0\}$.

24.4 Real and complex embeddings

Definition 24.5 Let K be a number field. An embedding σ of K in $\mathbb{C}$ over $\mathbb{Q}$ is called a *real embedding* if $\sigma(\alpha) \in \mathbb{R}$ for all $\alpha \in K$. Otherwise, it is called a *complex embedding*.

Proposition 24.8 Let K be a number field. Let $\theta \in K$ be such that $K = \mathbb{Q}(\theta)$. Then an embedding σ of K in $\mathbb{C}$ over $\mathbb{Q}$ is real if and only if $\sigma(\theta) \in \mathbb{R}$.

Proof. The necessity is trivial. To prove sufficiency, we note that every $\alpha \in K$ can be written as $\alpha = a_0 + a_1\theta + \cdots + a_{n-1}\theta^{n-1}$ with $a_i \in \mathbb{Q}$ for all i. Now, recalling that σ fixes $\mathbb{Q}$, we have $\sigma(\alpha) = a_0 + a_1\sigma(\theta) + \cdots + a_{n-1}\sigma(\theta)^{n-1} \in \mathbb{R}$ since $\sigma(\theta) \in \mathbb{R}$. ■

Lemma 24.9 Let $\alpha = a + b\sqrt{-1} \in \mathbb{C}$ be algebraic over $\mathbb{Q}$. Then its **complex conjugate** $\overline{\alpha} = a - b\sqrt{-1}$ is a conjugate of α over $\mathbb{Q}$.

Proof. Let $p(x) \in \mathbb{Q}[x]$ be the minimal polynomial of α over $\mathbb{Q}$. We have $p(\overline{\alpha}) = \overline{p(\alpha)} = \overline{0} = 0$, and hence obtain the desired claim. ■

For K a number field, if σ is an embedding of K in $\mathbb{C}$ over $\mathbb{Q}$, we define a homomorphism $\overline{\sigma} : K \to \mathbb{C}$ given by

$$\overline{\sigma}(\alpha) = \overline{\sigma(\alpha)},$$

the **complex conjugate** of $\sigma(\alpha)$ for all $\alpha \in K$. Let $\theta \in K$ be such that $K = \mathbb{Q}(\theta)$. Then by Theorem 22.2 and Lemma 24.9, the three numbers θ, $\sigma(\theta)$ and $\overline{\sigma}(\theta)$ are conjugates of each other over $\mathbb{Q}$, and hence $\overline{\sigma}$ is also an embedding of K in $\mathbb{C}$ over $\mathbb{Q}$.

Definition 24.6 Let K be a number field. For an embedding σ of K in $\mathbb{C}$ over $\mathbb{Q}$, the embedding $\overline{\sigma} : K \to \mathbb{C}$ given by

$$\overline{\sigma}(\alpha) = \overline{\sigma(\alpha)}$$

for all $\alpha \in K$ is called the *complex conjugate* of σ.

Note that the two embeddings σ and $\overline{\sigma}$ are complex conjugates of one another. In particular, if σ is real, then $\overline{\sigma} = \sigma$, and if σ is complex, then $\overline{\sigma} \neq \sigma$. It turns out that there are an even number of complex embeddings of K in $\mathbb{C}$ over $\mathbb{Q}$. Assuming that there are r_1 real and $2r_2$ complex embeddings of K in $\mathbb{C}$ over $\mathbb{Q}$ so that $r_1 + 2r_2 = [K : \mathbb{Q}]$, we may group these embeddings as follows:

$$\{\sigma_1\}, \ldots, \{\sigma_{r_1}\}, \{\sigma_{r_1+1}, \sigma_{r_1+2}\}, \ldots, \{\sigma_{r_1+2r_2-1}, \sigma_{r_1+2r_2}\}, \tag{24.5}$$

where $\sigma_1, \ldots, \sigma_{r_1}$ are real and $\sigma_{r_1+1}, \ldots, \sigma_{r_1+2r_2}$ are complex with σ_{r_1+2k-1} and σ_{r_1+2k} complex conjugates of one another for $1 \leq k \leq r_2$.

Theorem 24.10 Let K be a number field with exactly $2r_2$ complex embedding in $\mathbb{C}$ over $\mathbb{Q}$. The sign of the discriminant $d_K \in \mathbb{Z}\backslash\{0\}$ of K is $(-1)^{r_2}$.

Proof. Let the $n = r_1 + 2r_2 = [K : \mathbb{Q}]$ embeddings of K in $\mathbb{C}$ over $\mathbb{Q}$ be labeled as in (24.5). Let $\{\omega_1, \ldots, \omega_n\}$ be an integral basis for $\mathscr{O}_K$. Then $d_K = \left(\det\left(\sigma_i(\omega_j)\right)\right)^2$. Note also that

$$\overline{\det\left(\sigma_i(\omega_j)\right)} = \det\left(\overline{\sigma_i(\omega_j)}\right) = \det\left(\overline{\sigma_i}(\omega_j)\right) = (-1)^{r_2}\det\left(\sigma_i(\omega_j)\right).$$

The last equality holds as there are r_2 row exchanges between the matrices $\left(\overline{\sigma_i}(\omega_j)\right)$ and $\left(\sigma_i(\omega_j)\right)$, namely, the rows regarding σ_{r_1+2k-1} and σ_{r_1+2k} are exchanged for $1 \leq k \leq r_2$. Finally,

$$d_K = \left(\det\left(\sigma_i(\omega_j)\right)\right)^2 = (-1)^{r_2}\det\left(\sigma_i(\omega_j)\right)\overline{\det\left(\sigma_i(\omega_j)\right)} = (-1)^{r_2}\left|\det\left(\sigma_i(\omega_j)\right)\right|^2.$$

Noting that $\left|\det\left(\sigma_i(\omega_j)\right)\right|^2 \in \mathbb{Z}_{>0}$ gives the desired result. ■

25. Factorization in a ring of integers

25.1 Divisibility and congruences

Now we briefly discuss some basic factorization properties for a ring of algebraic integers. Throughout, let K be a number field and let $\mathscr{O}_K$ be its ring of integers. We first define divisibility and congruences in $\mathscr{O}_K$ in analogy to what we have done in $\mathbb{Z}$.

Definition 25.1 Let $\alpha, \beta \in \mathscr{O}_K$. We say β *divides* α, or α *is divisible by* β, denoted by $\beta \mid \alpha$, if there is an element $\xi \in \mathscr{O}_K$ such that $\alpha = \beta\xi$.

Definition 25.2 Let $\mu \in \mathscr{O}_K$ with $\mu \neq 0$. For any $\alpha, \beta \in \mathscr{O}_K$, we say that α *is congruent to* β *modulo* μ if $\mu \mid (\alpha - \beta)$. We write $\alpha \equiv \beta \pmod{\mu}$. If $\mu \nmid (\alpha - \beta)$, we write $\alpha \not\equiv \beta \pmod{\mu}$.

The following properties are immediate.

Theorem 25.1 Assume that all variables in this theorem are in $\mathscr{O}_K$.

(i) If $\alpha \mid \beta$, then $\alpha \mid \beta\gamma$;
(ii) If $\alpha \mid \beta$ and $\beta \mid \gamma$, then $\alpha \mid \gamma$;
(iii) If $\alpha \mid \beta$, then $\alpha\gamma \mid \beta\gamma$;
(iv) If $\alpha \mid \beta_i$ for $i = 1, \ldots, r$, then $\alpha \mid (\nu_1\beta_1 + \cdots + \nu_r\beta_r)$.

Theorem 25.2 Assume that all variables in this theorem are in $\mathscr{O}_K$ with $\mu \neq 0$.

(i) $\alpha \equiv \alpha \pmod{\mu}$;
(ii) If $\alpha \equiv \beta \pmod{\mu}$, then $\beta \equiv \alpha \pmod{\mu}$;
(iii) If $\alpha \equiv \beta \pmod{\mu}$ and $\beta \equiv \gamma \pmod{\mu}$, then $\alpha \equiv \gamma \pmod{\mu}$;
(iv) If $\alpha_1 \equiv \beta_1 \pmod{\mu}$ and $\alpha_2 \equiv \beta_2 \pmod{\mu}$, then

$$\alpha_1 + \alpha_2 \equiv \beta_1 + \beta_2 \pmod{\mu},$$
$$\alpha_1\alpha_2 \equiv \beta_1\beta_2 \pmod{\mu};$$

(v) If $\alpha \equiv \beta \pmod{\mu}$, then for any positive integer k,

$$\alpha^k \equiv \beta^k \pmod{\mu};$$

(vi) If $f(x_1,x_2,\dots)$ is a multivariate polynomial with coefficients in $\mathscr{O}_K$, and $\alpha_1 \equiv \beta_1 \pmod{\mu}$, $\alpha_2 \equiv \beta_2 \pmod{\mu}$, ..., then

$$f(\alpha_1,\alpha_2,\dots) \equiv f(\beta_1,\beta_2,\dots) \pmod{\mu}.$$

25.2 Units, irreducible elements and prime elements

Recall from Theorem 23.15 that $N_{K/\mathbb{Q}}(\alpha) \in \mathbb{Z}$ for every $\alpha \in \mathscr{O}_K$.

Proposition 25.3 Let $\alpha,\beta \in \mathscr{O}_K$. If $\beta \mid \alpha$, then $N_{K/\mathbb{Q}}(\beta)$ divides $N_{K/\mathbb{Q}}(\alpha)$ in $\mathbb{Z}$.

Proof. Let us write $\alpha = \beta\gamma$ for some $\gamma \in \mathscr{O}_K$. By Proposition 23.13, we have $N_{K/\mathbb{Q}}(\alpha) = N_{K/\mathbb{Q}}(\beta)N_{K/\mathbb{Q}}(\gamma)$. Noting that $N_{K/\mathbb{Q}}(\alpha)$, $N_{K/\mathbb{Q}}(\beta)$ and $N_{K/\mathbb{Q}}(\gamma)$ are in $\mathbb{Z}$, we are done. ■

Definition 25.3 An element $u \in \mathscr{O}_K$ is called a *unit* if there is an element $v \in \mathscr{O}_K$ such that $uv = 1$.

Theorem 25.4 An element $u \in \mathscr{O}_K$ is a unit if and only if $N_{K/\mathbb{Q}}(u) = \pm 1$.

Proof. We start with necessity. If u is a unit of $\mathscr{O}_K$, then there is an element $v \in \mathscr{O}_K$ such that $uv = 1$. Now, $N_{K/\mathbb{Q}}(u)N_{K/\mathbb{Q}}(v) = N_{K/\mathbb{Q}}(uv) = N_{K/\mathbb{Q}}(1) = 1$. Since $N_{K/\mathbb{Q}}(u)$ and $N_{K/\mathbb{Q}}(v)$ are in $\mathbb{Z}$, we must have $N_{K/\mathbb{Q}}(u) = \pm 1$.

We then prove sufficiency. Suppose $N_{K/\mathbb{Q}}(u) = \pm 1$. We know from Theorem 23.15 that $N_{K/\mathbb{Q}}(u) = N_{\mathbb{Q}(u)/\mathbb{Q}}(u)^{[K:\mathbb{Q}(u)]}$. Since $N_{\mathbb{Q}(u)/\mathbb{Q}}(u) \in \mathbb{Z}$, we have $N_{\mathbb{Q}(u)/\mathbb{Q}}(u) = \pm 1$. Let $u = u_1, u_2, \dots, u_r$ be the conjugates of u over $\mathbb{Q}$. By Corollary 23.12, they are integral over $\mathbb{Z}$. Now we choose $v = N_{\mathbb{Q}(u)/\mathbb{Q}}(u) \cdot u_2 \cdots u_r$, which is also integral over $\mathbb{Z}$. Then

$$uv = u_1 \cdot \big(N_{\mathbb{Q}(u)/\mathbb{Q}}(u) \cdot u_2 \cdots u_r\big) = N_{\mathbb{Q}(u)/\mathbb{Q}}(u) \cdot u_1 u_2 \cdots u_r = N_{\mathbb{Q}(u)/\mathbb{Q}}(u)^2 = (\pm 1)^2 = 1.$$

Finally, noting that $v = u^{-1} \in K$, we have $v \in \mathscr{O}_K$. Hence, u is a unit of $\mathscr{O}_K$. ■

Definition 25.4 Let $\alpha \in \mathscr{O}_K$. An element $\beta \in \mathscr{O}_K$ is called an *associate of α in $\mathscr{O}_K$* if $\beta = u\alpha$ with u a unit of $\mathscr{O}_K$.

Definition 25.5 A nonzero nonunit element $\alpha \in \mathscr{O}_K$ is called *irreducible* if $\beta \in \mathscr{O}_K$ dividing α implies that either β is a unit of $\mathscr{O}_K$, or β is an associate of α in $\mathscr{O}_K$.

Proposition 25.5 Let $\alpha \in \mathscr{O}_K$. If $|N_{K/\mathbb{Q}}(\alpha)|$ is a prime in $\mathbb{Z}$, α is irreducible in $\mathscr{O}_K$.

Proof. Since $|N_{K/\mathbb{Q}}(\alpha)|$ is a prime integer, α is neither zero nor a unit of $\mathscr{O}_K$. If α is not irreducible, we may write $\alpha = \beta\gamma$ with $\beta,\gamma \in \mathscr{O}_K$ nonzero and nonunit. By Theorem 25.4, we know that $|N_{K/\mathbb{Q}}(\beta)|$ and $|N_{K/\mathbb{Q}}(\gamma)|$ are integers greater than 1. However, $|N_{K/\mathbb{Q}}(\alpha)| = |N_{K/\mathbb{Q}}(\beta\gamma)| = |N_{K/\mathbb{Q}}(\beta)||N_{K/\mathbb{Q}}(\gamma)|$ is assumed to be a prime integer, and we are led to a contradiction. ■

Theorem 25.6 Every nonzero nonunit element in $\mathscr{O}_K$ is a finite product of irreducible elements in $\mathscr{O}_K$.

Proof. Let $\alpha \in \mathscr{O}_K$ be nonzero and nonunit. Then $|N_{K/\mathbb{Q}}(\alpha)| \geq 2$ is in $\mathbb{Z}$. We apply induction on $|N_{K/\mathbb{Q}}(\alpha)|$. If $|N_{K/\mathbb{Q}}(\alpha)| = 2$, then α itself is irreducible by Proposition 25.5. Now for $|N_{K/\mathbb{Q}}(\alpha)| \geq 2$, if α is irreducible, then we are done. If not, we may find nonzero nonunit elements $\beta,\gamma \in \mathscr{O}_K$ such that $\alpha = \beta\gamma$. Noting that $N_{K/\mathbb{Q}}(\alpha) = N_{K/\mathbb{Q}}(\beta)N_{K/\mathbb{Q}}(\gamma)$, we have $2 \le |N_{K/\mathbb{Q}}(\beta)|, |N_{K/\mathbb{Q}}(\gamma)| < |N_{K/\mathbb{Q}}(\alpha)|$. By the inductive hypothesis, β and γ are finite products of irreducible elements in $\mathscr{O}_K$, and so is $\alpha = \beta\gamma$. ■

Now a natural question is *whether the finite product factorization in Theorem 25.6 unique, up to reordering and associates?* Unfortunately, the answer is *negative* for generic $\mathscr{O}_K$. An instance will be presented in Sect. 26.2. Recalling that one crucial ingredient in our proof of the Fundamental Theorem of Arithmetic in $\mathbb{Z}$ is Corollary 2.7. In analogy, we shall define *prime elements* in $\mathscr{O}_K$.

Definition 25.6 A nonzero nonunit element $\pi \in \mathscr{O}_K$ is called a *prime element* if the *Euclid Condition* holds: $\pi \mid \alpha\beta$ implies that $\pi \mid \alpha$ or $\pi \mid \beta$ for any $\alpha,\beta \in \mathscr{O}_K$.

R Prime elements in $\mathscr{O}_{\mathbb{Q}} = \mathbb{Z}$ are exactly prime integers and their additive inverse. In particular, we shall call prime integers *rational primes* to avoid ambiguity.

Proposition 25.7 Every prime element in $\mathscr{O}_K$ is irreducible.

Proof. Let π be a prime element in $\mathscr{O}_K$ and suppose that $\pi = \alpha\beta$ is a factorization of π with $\alpha,\beta \in \mathscr{O}_K$. Then π divides one of α and β, say, $\pi \mid \alpha$ by definition. Now writing $\alpha = \pi\gamma$ with $\gamma \in \mathscr{O}_K$ gives $\pi = \alpha\beta = (\pi\gamma)\beta = \pi(\beta\gamma)$, and hence $\beta\gamma = 1$. So β is a unit, thereby implying the irreducibility of π. ■

Theorem 25.8 Let π be a prime element in $\mathscr{O}_K$. Then π divides a unique rational prime.

Proof. Assume that $\pi = \pi_1, \pi_2, \ldots, \pi_r$ are the conjugates of π over $\mathbb{Q}$. By Corollary 23.12, they are integral over $\mathbb{Z}$. Let $\pi' = \pi_2 \cdots \pi_r$, which is also integral over $\mathbb{Z}$. Note that $\pi\pi' = \pi_1\pi_2\cdots\pi_r = N_{\mathbb{Q}(\pi)/\mathbb{Q}}(\pi) \in \mathbb{Z}$. We have $\pi' = \pi^{-1}N_{\mathbb{Q}(\pi)/\mathbb{Q}}(\pi) \in K$, yielding that $\pi' \in \mathscr{O}_K$. Hence, π divides $|N_{\mathbb{Q}(\pi)/\mathbb{Q}}(\pi)| \in \mathbb{Z}_{>0}$. Recall that $\pi \nmid 1$ as π is nonunit. Let $p \ge 2$ be the smallest positive integer such that $\pi \mid p$. Then p must be a rational prime since π is prime in $\mathscr{O}_K$. Otherwise, if $p = p_1p_2$ with $2 \le p_1, p_2 < p$, then π divides one of p_1 and p_2, thereby violating the minimality of p. Finally, if π divides two different rational primes p and p', then by Theorem 2.5, there exist $a, a' \in \mathbb{Z}$ such that $ap + a'p' = 1$. Hence, $\pi \mid (ap + a'p') = 1$, which gives a contradiction. ■

Now there is a dilemma: if we factor elements in a ring $\mathscr{O}_K$ of integers with irreducible elements, then such factorizations might not be unique; if we factor elements in $\mathscr{O}_K$ with prime elements, then such factorizations might not exist. Both issues may happen for the same $\mathscr{O}_K$; see Sect. 26.2. Fortunately, we are not at a dead end. With prime elements in $\mathscr{O}_K$, it is still able to recover unique factorization by passing to ideals. Such a brilliant idea is due to the German mathematician Richard Dedekind. However, this highly algebraic topic will not be covered in the current series of notes.

25.3 Fundamental theorem of arithmetic revisited

We shall say more about rings of integers in which the Fundamental Theorem of Arithmetic remains valid.

Definition 25.7 Let K be a number field and let $\mathscr{O}_K$ be its ring of integers. We say $\mathscr{O}_K$ is a *unique factorization domain* if every nonzero nonunit element in $\mathscr{O}_K$ has a unique (up to reordering and associates) representation as a finite product of irreducible elements in $\mathscr{O}_K$. This property is called the *Fundamental Theorem of Arithmetic in* $\mathscr{O}_K$.

Recall from Theorem 25.6 that every nonzero nonunit element in $\mathscr{O}_K$ is a finite product of irreducible elements in $\mathscr{O}_K$. Also, Proposition 25.7 tells us that every prime element in $\mathscr{O}_K$ is irreducible, but the converse may be false. Interestingly, we can determine if $\mathscr{O}_K$ is a unique factorization domain from this perspective.

Theorem 25.9 Let K be a number field. Then $\mathscr{O}_K$ is a unique factorization domain if and only if every irreducible element in $\mathscr{O}_K$ is prime.

Proof. The sufficiency can be proved in a similar way to that for the Fundamental Theorem of Arithmetic in $\mathbb{Z}$. Recall that if π is a prime element in $\mathscr{O}_K$, then $\pi \mid \alpha$ or $\pi \mid \beta$ for any $\alpha,\beta \in \mathscr{O}_K$ whenever $\pi \mid \alpha\beta$. Now, if $\pi \mid \pi_1 \cdots \pi_k$ with $\pi_1,\ldots,\pi_k$ prime in $\mathscr{O}_K$, then π is an associate of π_j for at least one j. Assume that a nonzero nonunit element $\alpha \in \mathscr{O}_K$ has factorizations

$$\alpha = \pi_1 \cdots \pi_k = \pi_1' \cdots \pi_\ell'$$

with π_i and π_j' irreducible, and hence prime by assumption, in $\mathscr{O}_K$. Then without loss of generality, we have $\pi_1 = u\pi_1'$ with u a unit of $\mathscr{O}_K$. Thus, $u\pi_2 \cdots \pi_k = \pi_2' \cdots \pi_\ell'$. We may tacitly rename $u\pi_2$ by π_2 as they are associates of one another. In other words, we get $\pi_2 \cdots \pi_k = \pi_2' \cdots \pi_\ell'$. Repeating this process implies the unique factorization.

We then prove the necessity. Let ξ be an irreducible element in a unique factorization domain $\mathscr{O}_K$. Suppose $\xi \mid \alpha\beta$ where $\alpha,\beta \in \mathscr{O}_K$. If one of α,β is 0 or a unit, then it is plain that $\xi \mid \alpha$ or $\xi \mid \beta$. Now assume that both α,β are nonzero and nonunit. We factor $\alpha = \alpha_1 \cdots \alpha_r$ and $\beta = \beta_1 \cdots \beta_s$ with α_i and β_j irreducible. Recall that ξ is irreducible and that $\xi \mid \alpha\beta$. By the uniqueness of factorization of $\alpha\beta$, it is known that ξ is associated with some α_i or β_j, thereby implying that $\xi \mid \alpha$ or $\xi \mid \beta$. So ξ is a prime element in $\mathscr{O}_K$. ■

25.4 Norm-Euclidean number fields

Now we are facing the demand of examining the equivalence between irreducible elements and primes in a given $\mathscr{O}_K$. Let us recall that in our proof of such an equivalence in $\mathbb{Z}$, namely, Corollary 2.7, a key ingredient is the greatest common divisor, which comes from the Euclidean Algorithm. Further, the Euclidean Algorithm is built on the Division Algorithm: for any integers $a,b \in \mathbb{Z} = \mathscr{O}_\mathbb{Q}$ with $b \neq 0$, there are integers $q,r \in \mathbb{Z}$ such that

$$a = qb + r, \qquad |r| < |b|. \tag{25.1}$$

We also note that for any $a \in \mathbb{Q}$, we have $N_{\mathbb{Q}/\mathbb{Q}}(a) = a$. The above discussions suggest the following analogy.

Definition 25.8 Let K be a number field. We say K is *Norm-Euclidean* if for any $\alpha,\beta \in \mathscr{O}_K$ with $\beta \neq 0$, there are $\eta,\rho \in \mathscr{O}_K$ such that

$$\alpha = \eta\beta + \rho, \qquad |N_{K/\mathbb{Q}}(\rho)| < |N_{K/\mathbb{Q}}(\beta)|. \tag{25.2}$$

We may also transplant the definition of the greatest common divisor.

Definition 25.9 Let K be a norm-Euclidean number field. Let $\alpha,\beta \in \mathscr{O}_K$, not both zero. There exists a unique algebraic integer $\delta \in \mathscr{O}_K$, up to associates, such that δ divides both α and β, and such that if $\delta' \in \mathscr{O}_K$ divides α and β, then $\delta' \mid \delta$. This algebraic integer $\delta \in \mathscr{O}_K$ is called the *greatest common divisor* of α and β, denoted by $\delta = (\alpha,\beta)$.

To get this algebraic integer $\delta \in \mathscr{O}_K$, we shall still use the Euclidean Algorithm as follows. Without loss of generality, we assume that $|N_{K/\mathbb{Q}}(\alpha)| \geq |N_{K/\mathbb{Q}}(\beta)|$ and $\beta \neq 0$. Recall that for $\rho \in \mathscr{O}_K$, we have $N_{K/\mathbb{Q}}(\rho) \in \mathbb{Z}$ while $N_{K/\mathbb{Q}}(\rho) = 0$ if and only if $\rho = 0$. Let us put $\rho_{-1} = \alpha$ and $\rho_0 = \beta$. Now we iteratively write

$$\begin{aligned}
\rho_{-1} &= \eta_1\rho_0 + \rho_1, & 0 < |N_{K/\mathbb{Q}}(\rho_1)| < |N_{K/\mathbb{Q}}(\rho_0)|;\\
\rho_0 &= \eta_2\rho_1 + \rho_2, & 0 < |N_{K/\mathbb{Q}}(\rho_2)| < |N_{K/\mathbb{Q}}(\rho_1)|\\
\rho_1 &= \eta_3\rho_2 + \rho_3, & 0 < |N_{K/\mathbb{Q}}(\rho_3)| < |N_{K/\mathbb{Q}}(\rho_2)|;\\
&\cdots\\
\rho_{k-2} &= \eta_k\rho_{k-1} + \rho_k, & 0 < |N_{K/\mathbb{Q}}(\rho_k)| < |N_{K/\mathbb{Q}}(\rho_{k-1})|;\\
\rho_{k-1} &= \eta_{k+1}\rho_k + 0.
\end{aligned}$$

Then $\delta = \rho_k$ is as required.

We establish a Bézout-type identity analogous to Theorem 2.5.

Theorem 25.10 (Bézout's Identity for Norm-Euclidean Number Fields). Let K be a norm-Euclidean number field. Let $\alpha, \beta \in \mathscr{O}_K$, not both zero, and denote $\delta = (\alpha, \beta)$. Then there exist $\mu, \nu \in \mathscr{O}_K$ such that $\delta = \alpha\mu + \beta\nu$.

Proof. We only need the fact that the set $S = \{\alpha\mu + \beta\nu : \mu, \nu \in \mathscr{O}_K\}$ is closed under addition and scalar multiplication (of elements in $\mathscr{O}_K$). From the above Euclidean Algorithm, we iteratively have $\rho_1 \in S$, $\rho_2 \in S$, ..., and finally, $\delta = \rho_k \in S$. ■

Theorem 25.11 Let K be a norm-Euclidean number field. Then every irreducible element in $\mathscr{O}_K$ is prime. Consequently, $\mathscr{O}_K$ is a unique factorization domain.

Proof. We shall show that for any irreducible $\pi \in \mathscr{O}_K$, if $\pi \mid \alpha\beta$ with $\alpha, \beta \in \mathscr{O}_K$, then $\pi \mid \alpha$ or $\pi \mid \beta$. If $\pi \mid \alpha$, then we are done. If $\pi \nmid \alpha$, then $(\pi, \alpha) = 1$ since π is irreducible. By Theorem 25.10, we choose elements $\mu, \nu \in \mathscr{O}_K$ such that $1 = \pi\mu + \alpha\nu$. Then

$$\beta = \beta \cdot 1 = \beta(\pi\mu + \alpha\nu) = \pi \cdot (\beta\mu) + (\alpha\beta) \cdot \nu.$$

It follows that $\pi \mid \beta$. Hence, π is prime. Finally, we know from Theorem 25.9 that $\mathscr{O}_K$ is a unique factorization domain. ■

R We should point out that there *do* exist non-norm-Euclidean number fields K with $\mathscr{O}_K$ a unique factorization domain.

Finally, we note that the Division Algorithm (25.1) is equivalent to the claim that for all $x \in \mathbb{Q}$, there is an integer $n \in \mathbb{Z}$ such that

$$|x - n| < 1.$$

We have a parallel result for norm-Euclidean number fields.

Theorem 25.12 Let K be a number field. Then K is norm-Euclidean if and only if for all $\xi \in K$, there is an integer $\eta \in \mathscr{O}_K$ such that

$$\big|N_{K/\mathbb{Q}}(\xi - \eta)\big| < 1. \tag{25.3}$$

Proof. We start with necessity. Assume that K is norm-Euclidean. It is known from Theorem 23.9 that there is a nonzero integer $n \in \mathbb{Z}$ such that $n\xi \in \mathscr{O}_K$. Now since K is norm-Euclidean, we choose $\eta, \rho \in \mathscr{O}_K$ such that

$$n\xi = \eta n + \rho, \qquad |N_{K/\mathbb{Q}}(\rho)| < |N_{K/\mathbb{Q}}(n)|.$$

Thus,

$$\left|N_{K/\mathbb{Q}}(\xi-\eta)\right| = \left|N_{K/\mathbb{Q}}(\tfrac{\rho}{n})\right| < 1.$$

We then prove sufficiency. Assume the condition (25.3). Let $\alpha, \beta \in \mathscr{O}_K$ be arbitrary with $\beta \neq 0$. We choose $\eta \in \mathscr{O}_K$ such that $|N_{K/\mathbb{Q}}(\kappa)| < 1$ where $\kappa = \frac{\alpha}{\beta} - \eta$. Now, $\alpha = \eta\beta + \kappa\beta$. Finally, we note that

$$\left|N_{K/\mathbb{Q}}(\kappa\beta)\right| = \left|N_{K/\mathbb{Q}}(\kappa)\right| \left|N_{K/\mathbb{Q}}(\beta)\right| < \left|N_{K/\mathbb{Q}}(\beta)\right|,$$

as required. ■

26. Quadratic fields

26.1 Quadratic fields

Now we shall use quadratic fields as concrete examples to illustrate ideas in the previous lectures.

Definition 26.1 A *quadratic field* is a number field of degree 2 over $\mathbb{Q}$.

Recall that a nonzero integer d is called squarefree if no integer squares other than 1 divide d.

Theorem 26.1 Let K be a quadratic field. Then there exists a unique squarefree integer $d \neq 1$ such that $K = \mathbb{Q}(\sqrt{d})$.

Proof. Suppose that $\theta \in K$ is such that $K = \mathbb{Q}(\theta)$. Then the minimal polynomial $p(x)$ of θ over $\mathbb{Q}$ is of degree 2, say $p(x) = x^2 + a_1x + a_0$ with $a_0, a_1 \in \mathbb{Q}$. Solving $p(x) = 0$ gives two solutions $\theta_{1,2} = \frac{-a_1 \pm \sqrt{\Delta}}{2}$ where $\Delta = a_1^2 - 4a_0$; we shall require that $\sqrt{\Delta}$ is not rational to ensure that $p(x)$ is irreducible over $\mathbb{Q}$. Now, $K = \mathbb{Q}(\theta) = \mathbb{Q}(\sqrt{\Delta})$. Noting that $a_1^2 - 4a_0$ is in $\mathbb{Q}$ and is nonzero, we write $a_1^2 - 4a_0 = \frac{s}{t}$ with $s \neq 0$ and $t > 0$ in $\mathbb{Z}$. Thus, $\sqrt{\Delta} = \sqrt{\frac{s}{t}} = \frac{\sqrt{st}}{t} = \frac{r\sqrt{d}}{t}$ where we further write $st = r^2d$ with d squarefree. In particular, $d \neq 1$ since $\sqrt{st}$ is not rational as assumed.

To show the uniqueness of d, it suffices to prove that if $d_1 \neq 1$ and $d_2 \neq 1$ are two distinct squarefree integers, then $\mathbb{Q}(\sqrt{d_1}) \neq \mathbb{Q}(\sqrt{d_2})$. Here, we only need to show that 1, $\sqrt{d_1}$ and $\sqrt{d_2}$ are linearly independent over $\mathbb{Q}$. If not, then we have $a\sqrt{d_1} + b\sqrt{d_2} = c$ with $a, b, c \in \mathbb{Q}$ not all zero. It is plain that 1 and $\sqrt{d}$ are linearly independent over $\mathbb{Q}$, where $d \neq 1$ is squarefree. This is because if $d < 0$ then $\sqrt{d}$ is purely imaginary and if $d > 1$ then we recall Example 19.7. Hence, $a, b \neq 0$. Squaring both sides of $a\sqrt{d_1} + b\sqrt{d_2} = c$ implies that $\sqrt{d_1d_2} \in \mathbb{Q}$. However, if d_1 and d_2 have different signs, then $\sqrt{d_1d_2}$ is purely imaginary, which is not in $\mathbb{Q}$. If d_1 and d_2 have the same sign, then we note that there is a prime p such that $p \mid d_1d_2$ and $p^2 \nmid d_1d_2$ for $d_1 \neq d_2$ are squarefree. Recalling Example 19.7 again implies that in this case $\sqrt{d_1d_2}$ is not in $\mathbb{Q}$. We are therefore led to a contradiction. ■

Definition 26.2 Let $d \neq 1$ be a squarefree integer in $\mathbb{Z}$. The quadratic field $\mathbb{Q}(\sqrt{d})$ is called *real* if $d > 0$, and *imaginary* if $d < 0$.

Let us collect some basic facts about quadratic fields.

Fact 26.2 Let $d \neq 1$ be a squarefree integer in $\mathbb{Z}$.

(i) $\mathbb{Q}(\sqrt{d}) = \mathbb{Q}[\sqrt{d}]$. Also, $\{1, \sqrt{d}\}$ forms a basis for $\mathbb{Q}(\sqrt{d})$ over $\mathbb{Q}$. In what follows, we write elements in $\mathbb{Q}(\sqrt{d})$ as $a + b\sqrt{d}$ with $a, b \in \mathbb{Q}$.

(ii) The minimal polynomial of $a + b\sqrt{d}$ over $\mathbb{Q}$ is $p(x) = x^2 - 2ax + (a^2 - db^2)$. Futher, $a + b\sqrt{d}$ has two conjugates over $\mathbb{Q}$, namely, $a + b\sqrt{d}$ and $a - b\sqrt{d}$.

(iii) $\mathrm{Tr}_{\mathbb{Q}(\sqrt{d})/\mathbb{Q}}(a + b\sqrt{d}) = 2a \in \mathbb{Q}$.

(iv) $N_{\mathbb{Q}(\sqrt{d})/\mathbb{Q}}(a + b\sqrt{d}) = a^2 - db^2 \in \mathbb{Q}$. In particular, $N_{\mathbb{Q}(\sqrt{d})/\mathbb{Q}}(a + b\sqrt{d}) \geq 0$ if $d < 0$.

(v) There are exactly two embeddings σ_1 and σ_2 of $\mathbb{Q}(\sqrt{d})$ in $\mathbb{C}$ over $\mathbb{Q}$ given by $\sigma_1(a + b\sqrt{d}) = a + b\sqrt{d}$ and $\sigma_2(a + b\sqrt{d}) = a - b\sqrt{d}$ for all $a + b\sqrt{d} \in \mathbb{Q}(\sqrt{d})$. If $d > 0$, then both σ_1 and σ_2 are real; if $d < 0$, then both σ_1 and σ_2 are complex.

(vi) The quadratic extension $\mathbb{Q}(\sqrt{d})/\mathbb{Q}$ is Galois.

We then study the ring $\mathscr{O}_{\mathbb{Q}(\sqrt{d})}$ of integers of $\mathbb{Q}(\sqrt{d})$.

Theorem 26.3 Let $d \neq 1$ be a squarefree integer in $\mathbb{Z}$.

(i) $\mathscr{O}_{\mathbb{Q}(\sqrt{d})} = \begin{cases} \mathbb{Z}[\sqrt{d}], & \text{if } d \equiv 2 \text{ or } 3 \pmod 4, \\ \mathbb{Z}[\frac{1+\sqrt{d}}{2}], & \text{if } d \equiv 1 \pmod 4. \end{cases}$

(ii) If $d \equiv 2$ or $3 \pmod 4$, then $\{1, \sqrt{d}\}$ forms an integral basis for $\mathscr{O}_{\mathbb{Q}(\sqrt{d})}$; if $d \equiv 1 \pmod 4$, then $\{1, \frac{1+\sqrt{d}}{2}\}$ forms an integral basis for $\mathscr{O}_{\mathbb{Q}(\sqrt{d})}$.

(iii) $d_{\mathbb{Q}(\sqrt{d})} = \begin{cases} 4d, & \text{if } d \equiv 2 \text{ or } 3 \pmod 4, \\ d, & \text{if } d \equiv 1 \pmod 4. \end{cases}$

Proof. **(i).** We shall prove that

$$\mathscr{O}_{\mathbb{Q}(\sqrt{d})} = \begin{cases} \{a + b\sqrt{d} : a, b \in \mathbb{Z}\}, & \text{if } d \equiv 2 \text{ or } 3 \pmod 4, \\ \{\frac{a+b\sqrt{d}}{2} : a, b \in \mathbb{Z} \text{ and } a \equiv b \bmod 2\}, & \text{if } d \equiv 1 \pmod 4. \end{cases}$$

Consider $a + b\sqrt{d} \in \mathbb{Q}(\sqrt{d})$ where $a, b \in \mathbb{Q}$. We know from Corollary 23.12 that $a + b\sqrt{d} \in \mathscr{O}_{\mathbb{Q}(\sqrt{d})}$ if and only if its minimal polynomial over $\mathbb{Q}$,

$$p(x) = x^2 - 2ax + (a^2 - db^2),$$

is in $\mathbb{Z}[x]$, i.e. $2a$ and $a^2 - db^2$ are simultaneously in $\mathbb{Z}$. If $a \in \mathbb{Z}$, then $db^2 \in \mathbb{Z}$ implies that $b \in \mathbb{Z}$ as d is squarefree. The other possibility is that $a \in \mathbb{Z} + \frac{1}{2}$ so that $db^2 \in \mathbb{Z} + \frac{1}{4}$. Since d is squarefree, we must have $b \in \mathbb{Z} + \frac{1}{2}$, which further requires that $d \equiv 1 \pmod 4$.

(ii). This is a direct consequence of Part (i).

(iii). If $d \equiv 2$ or $3 \pmod 4$, we have

$$d_{\mathbb{Q}(\sqrt{d})} = \mathrm{disc}(1, \sqrt{d}) = \left(\det \begin{pmatrix} 1 & \sqrt{d} \\ 1 & -\sqrt{d} \end{pmatrix}\right)^2 = 4d.$$

If $d \equiv 1 \pmod 4$, we have

$$d_{\mathbb{Q}(\sqrt{d})} = \mathrm{disc}(1, \tfrac{1+\sqrt{d}}{2}) = \left(\det \begin{pmatrix} 1 & \frac{1+\sqrt{d}}{2} \\ 1 & \frac{1-\sqrt{d}}{2} \end{pmatrix}\right)^2 = d,$$

as desired. ■

Finally, we determine the units of the ring of integers of a quadratic field. Let us begin with the imaginary quadratic fields.

Theorem 26.4 (Units of Imaginary Quadratic Fields). Let $d<0$ be a squarefree integer in $\mathbb{Z}$. Then the units of $\mathscr{O}_{\mathbb{Q}(\sqrt{d})}$ are

$$\begin{cases} \pm 1, \pm\sqrt{-1}, & \text{if } d=-1,\\ \pm 1, \frac{\pm 1 \pm \sqrt{-3}}{2}, & \text{if } d=-3,\\ \pm 1, & \text{otherwise.} \end{cases}$$

Proof. If $d \equiv 2$ or $3 \pmod 4$, elements in $\mathscr{O}_{\mathbb{Q}(\sqrt{d})}$ are of the form $a+b\sqrt{d}$ with $a,b\in\mathbb{Z}$. Also, if it is a unit, then by Theorem 25.4, $N_{\mathbb{Q}(\sqrt{d})/\mathbb{Q}}(a+b\sqrt{d}) = a^2-db^2 = \pm 1$. If $d=-1$, we have solutions $(a,b)=(\pm 1,0)$ and $(0,\pm 1)$; if $d<-1$, we have solutions $(a,b)=(\pm 1,0)$.

If $d\equiv 1 \pmod 4$, elements in $\mathscr{O}_{\mathbb{Q}(\sqrt{d})}$ are of the form $\frac{a+b\sqrt{d}}{2}$ with $a,b\in\mathbb{Z}$ and $a\equiv b \pmod 2$. If it is a unit, then $N_{\mathbb{Q}(\sqrt{d})/\mathbb{Q}}(\frac{a+b\sqrt{d}}{2}) = \frac{a^2-db^2}{4} = \pm 1$. If $d=-3$, we have solutions $(a,b)=(\pm 2,0)$ and $(\pm 1,\pm 1)$; if $d<-3$, we have solutions $(a,b)=(\pm 2,0)$. ■

For real quadratic fields, units have a very different character.

Units of Real Quadratic Fields Let $d>1$ be a squarefree integer in $\mathbb{Z}$. Then the units of $\mathscr{O}_{\mathbb{Q}(\sqrt{d})}$ are real. Further, there exists a unique unit $\varepsilon>1$ such that all units of $\mathscr{O}_{\mathbb{Q}(\sqrt{d})}$ are of the form $\pm\varepsilon^n$ with $n\in\mathbb{Z}$.

Its proof will be postponed until a later lecture in Sect. 30.4. Here we only prepare a lemma a future use.

Lemma 26.5 Let $d>1$ be a squarefree integer in $\mathbb{Z}$. Then the set of units of $\mathscr{O}_{\mathbb{Q}(\sqrt{d})}$ is given by

$$\left\{a+b\sqrt{d} : a,b\in \tfrac{1}{2}\mathbb{Z} \text{ and } a^2-db^2=\pm 1\right\}.$$

Proof. It is plain that all units of $\mathscr{O}_{\mathbb{Q}(\sqrt{d})}$ are in the given set. Now it suffices to show that all elements in this set are units of $\mathscr{O}_{\mathbb{Q}(\sqrt{d})}$, and in fact, it is enough to show that these elements are in $\mathscr{O}_{\mathbb{Q}(\sqrt{d})}$. Since $a^2-db^2=\pm 1$ while d is squarefree, we find that a and b are simultaneously in $\mathbb{Z}$ or simultaneously in $\mathbb{Z}+\frac{1}{2}$. Hence, the case where $d\equiv 1 \pmod 4$ is automatically proved. For $d\equiv 2,3 \pmod 4$, we shall show that the situation that a and b are simultaneously in $\mathbb{Z}+\frac{1}{2}$ will never happen. However, if this is the case, then there are $u,v\in\mathbb{Z}$ such that

$$\left(\frac{2u+1}{2}\right)^2 - d\left(\frac{2v+1}{2}\right)^2 = \pm 1,$$

that is,

$$(2u+1)^2-d(2v+1)^2=\pm 4.$$

But we have $(2u+1)^2-d(2v+1)^2 \equiv 1-d \not\equiv 0 \pmod 4$, thereby yielding a contradiction. ■

26.2 Quadratic field $\mathbb{Q}(\sqrt{-5})$

As we had promised in Sect. 25.2, here we shall use the field $\mathbb{Q}(\sqrt{-5})$ to illustrate number fields in which the Fundamental Theorem of Arithmetic is false.

Claim 26.6 Consider the quadratic field $\mathbb{Q}(\sqrt{-5})$ where $\mathscr{O}_{\mathbb{Q}(\sqrt{-5})} = \mathbb{Z}[\sqrt{-5}]$.

(i) The numbers 2, 3 and $1\pm\sqrt{-5}$ are irreducible in $\mathbb{Z}[\sqrt{-5}]$.

(ii) The number 6 can be factored in two different ways by irreducible elements in $\mathbb{Z}[\sqrt{-5}]$, namely, $6 = 2 \times 3 = (1+\sqrt{-5}) \times (1-\sqrt{-5})$. Here, 2 and 3 are not associates of $1 \pm \sqrt{-5}$.

(iii) The numbers 2, 3 and $1 \pm \sqrt{-5}$ are not prime in $\mathbb{Z}[\sqrt{-5}]$.

Proof. Note that the norms of 2, 3, $1+\sqrt{-5}$ and $1-\sqrt{-5}$ to $\mathbb{Q}$ are 4, 9, 6 and 6, respectively, and hence that 2 and 3 are not associates of $1 \pm \sqrt{-5}$.

If any of them is not irreducible, then $\mathbb{Z}[\sqrt{-5}]$ would contain an element $a+b\sqrt{-5}$ with $a,b \in \mathbb{Z}$ such that $N_{\mathbb{Q}(\sqrt{-5})/\mathbb{Q}}(a+b\sqrt{-5}) = a^2+5b^2 \in \{\pm 2, \pm 3\}$. However, we cannot find such a and b.

Finally, we recall that $2 \mid (1+\sqrt{-5})(1-\sqrt{-5})$. If 2 divides $1 \pm \sqrt{-5}$, so does the norm to $\mathbb{Q}$. But $4 \nmid 6$, and we have a contradiction. Hence, 2 is not prime in $\mathbb{Z}[\sqrt{-5}]$. Similarly, we can show that 3 and $1 \pm \sqrt{-5}$ are not prime in $\mathbb{Z}[\sqrt{-5}]$. ■

26.3 Norm-Euclidean imaginary quadratic number fields

Our next object is to determine all imaginary quadratic number fields that are norm-Euclidean.

Theorem 26.7 There are exactly 5 norm-Euclidean imaginary quadratic number fields, namely, $\mathbb{Q}(\sqrt{-1})$, $\mathbb{Q}(\sqrt{-2})$, $\mathbb{Q}(\sqrt{-3})$, $\mathbb{Q}(\sqrt{-7})$ and $\mathbb{Q}(\sqrt{-11})$.

Proof. We only need to consider $\mathbb{Q}(\sqrt{d})$ with $d<0$ a squarefree integer. Recall that for any $\alpha = a+b\sqrt{d} \in \mathbb{Q}(\sqrt{d})$ with $a,b \in \mathbb{Q}$, we have $N_{\mathbb{Q}(\sqrt{d})/\mathbb{Q}}(\alpha) = a^2-db^2$.

For convenience, we introduce two auxiliary functions on $\mathbb{R}$. First, we define $I(x) := \min\{|x-n| : n \in \mathbb{Z}\}$, namely, the minimal distance between $x \in \mathbb{R}$ and an integer in $\mathbb{Z}$. Next, we define $H(x) := \min\{|x-(n+\frac{1}{2})| : n \in \mathbb{Z}\}$, namely, the minimal distance between $x \in \mathbb{R}$ and a half-integer in $\mathbb{Z}+\frac{1}{2}$. It is plain that for all $x \in \mathbb{R}$, we have $0 \le I(x) \le \frac{1}{2}$ and $0 \le H(x) \le \frac{1}{2}$. Also, $I(x)+H(x) = \frac{1}{2}$.

(i). Assume that $d \equiv 2$ or $3 \pmod 4$. Then $\mathscr{O}_{\mathbb{Q}(\sqrt{d})} = \mathbb{Z}[\sqrt{d}]$. Now,

$$\min\left\{\left|N_{\mathbb{Q}(\sqrt{d})/\mathbb{Q}}(\alpha-\eta)\right| : \eta \in \mathscr{O}_{\mathbb{Q}(\sqrt{d})}\right\} = I(a)^2+I(b)^2|d|.$$

If $d=-1$ or -2, then the right-hand side of the above is at most $(\frac{1}{2})^2+(\frac{1}{2})^2|d| = \frac{1+|d|}{4} < 1$ for all $\alpha \in \mathbb{Q}(\sqrt{d})$. If $d \le -5$, then we choose $\alpha = \frac{1}{2}+\frac{1}{2}\sqrt{d}$ so that the right-hand side of the above becomes $(\frac{1}{2})^2+(\frac{1}{2})^2|d| = \frac{1+|d|}{4} \ge 1$. Applying Theorem 25.12 confirms that $\mathbb{Q}(\sqrt{-1})$ and $\mathbb{Q}(\sqrt{-2})$ are the only norm-Euclidean number fields in this case.

(ii). Assume that $d \equiv 1 \pmod 4$. Then $\mathscr{O}_{\mathbb{Q}(\sqrt{d})} = \mathbb{Z}[\frac{1+\sqrt{d}}{2}]$. Now,

$$\begin{aligned}\min\left\{\left|N_{\mathbb{Q}(\sqrt{d})/\mathbb{Q}}(\alpha-\eta)\right| : \eta \in \mathscr{O}_{\mathbb{Q}(\sqrt{d})}\right\} &= \min\left\{I(a)^2+I(b)^2|d|, H(a)^2+H(b)^2|d|\right\}\\ &= \min\left\{I(a)^2+I(b)^2|d|, (\tfrac{1}{2}-I(a))^2+(\tfrac{1}{2}-I(b))^2|d|\right\}.\end{aligned}$$

If $d=-3$, -7 or -11, then the right-hand side of the above is at most $(\frac{1}{2})^2+(\frac{1}{4})^2|d| = \frac{4+|d|}{16} < 1$ for all $\alpha \in \mathbb{Q}(\sqrt{d})$. If $d \le -15$, then we choose $\alpha = \frac{1}{4}+\frac{1}{4}\sqrt{d}$ so that the right-hand side of the above becomes $(\frac{1}{4})^2+(\frac{1}{4})^2|d| = \frac{1+|d|}{16} \ge 1$. By Theorem 25.12, $\mathbb{Q}(\sqrt{-3})$, $\mathbb{Q}(\sqrt{-7})$ and $\mathbb{Q}(\sqrt{-11})$ are the only norm-Euclidean number fields in this case. ■

R For norm-Euclidean real quadratic number fields, it is known that there are exactly 16 of them, namely, $\mathbb{Q}(\sqrt{d})$ with

$$d \in \{2,3,5,6,7,11,13,17,19,21,29,33,37,41,57,73\}.$$

This result is essentially due to Harold Chatland and Harold Davenport (*Canad. J. Math.* **2** (1950), 289–296). Also, there are exactly 4 non-norm-Euclidean imaginary quadratic number fields that are unique factorization domains, namely, $\mathbb{Q}(\sqrt{d})$ with

$$d \in \{-19,-43,-67,-163\}.$$

Hans Heilbronn and Edward Linfoot (*Quart. J. Math. Oxford Ser.* **5** (1934), 150–160 & 293–301) first proved that there would be at most one more instance other than the above choices. Harold Stark (*Michigan Math. J.* **14** (1967), 1–27) further removed this fabricated field. The case of non-norm-Euclidean real quadratic number fields is more intricate. One example is $\mathbb{Q}(\sqrt{14})$; see Malcolm Harper's Ph.D. thesis.

26.4 Quadratic field $\mathbb{Q}(\sqrt{-1})$

Let us adopt the conventional notation $i = \sqrt{-1}$. Recall that $\mathscr{O}_{\mathbb{Q}(\sqrt{-1})} = \mathbb{Z}[i]$.

Fact 26.8 $\mathbb{Z}[i]$ is a unique factorization domain, i.e. every nonzero nonunit element in $\mathbb{Z}[i]$ has a unique (up to reordering and associates) representation as a finite product of irreducible (or equivalently, prime) elements in $\mathbb{Z}[i]$.

Now it remains to determine all irreducible (or equivalently, prime) elements in $\mathbb{Z}[i]$. We know from Theorem 25.8 that prime elements in $\mathbb{Z}[i]$ are nonunit factors of rational primes p. Note that $N_{\mathbb{Q}(\sqrt{-1})/\mathbb{Q}}(p) = p^2$. Also, if a nonunit $\alpha = a+bi \in \mathbb{Z}[i]$ with $a,b \in \mathbb{Z}$ is such that $\alpha \mid p$ and that α is not an associate of p, then $N_{\mathbb{Q}(\sqrt{-1})/\mathbb{Q}}(\alpha) = p$, and hence,

$$p = a^2 + b^2. \tag{26.1}$$

For a given rational prime p, if such an α does not exist, i.e. (26.1) has no integer solution (a,b), then p itself is irreducible and hence prime in $\mathbb{Z}[i]$, and in this case we have irreducible elements in $\mathbb{Z}[i]$ given by p and its associates. If such an α exists, we assume that $\beta \in \mathbb{Z}[i]$ is such that $p = \alpha\beta$, and therefore that $N_{\mathbb{Q}(\sqrt{-1})/\mathbb{Q}}(\beta) = p$. Thus, α and β are prime (and hence irreducible) elements in $\mathbb{Z}[i]$ by Proposition 25.5, and the factorization of $p = \alpha\beta$ by irreducible elements in $\mathbb{Z}[i]$ is unique, up to reordering and associates. So in this case we have irreducible elements in $\mathbb{Z}[i]$ given by α, β and their associates, with duplicates removed.

Let us first recover a special case of Jacobi's two-square formula (12.5).

Theorem 26.9 Let $p \equiv 1 \pmod 4$ be a rational prime. Then the Diophantine equation

$$p = m^2 + n^2 \tag{26.2}$$

has exactly 8 solutions for $m,n \in \mathbb{Z}$.

Proof. It is known from Theorem 6.10 that $\left(\frac{-1}{p}\right) = 1$ for $p \equiv 1 \pmod 4$, and hence that there is an integer $x \in \mathbb{Z}$ such that $p \mid (x^2+1)$. Further, in $\mathbb{Z}[i]$, we have the factorization $x^2+1 = (x+i)(x-i)$. We claim that p is not prime (and hence not irreducible) in $\mathbb{Z}[i]$. If not, then $p \mid (x+i)$ or $p \mid (x-i)$. But this is impossible as $\frac{x \pm i}{p} \notin \mathbb{Z}[i]$. Thus, we may uniquely factor $p = \alpha\beta$ with $\alpha,\beta \in \mathbb{Z}[i]$ nonunit. In particular, if $\alpha = a+bi$, then $\beta = a-bi$, and it is plain that $|a| \neq |b|$ and that a, b are nonzero. Recalling from Theorem 26.4 that the units of $\mathbb{Z}[i]$ are ± 1 and $\pm i$, we see that β is not an associate of α. It turns out that (26.2) has exactly 8 solutions, determined by the 4 associates of α and the 4 associates of β. ■

Theorem 26.10 The irreducible (or equivalently, prime) elements in $\mathbb{Z}[i]$ are:

(i) $1+i$ and its associates;

(ii) rational primes p and their associates for $p \equiv 3 \pmod 4$;

(iii) nonunit and nonassociate factors $a+bi$ of rational primes p for $p \equiv 1 \pmod 4$, i.e. $a, b \in \mathbb{Z}$ are such that $p = a^2 + b^2$.

Proof. There are three cases: **(i).** If $p = 2$, then we have the factorization $2 = (1+i)(1-i)$ where $1+i$ and $1-i$ are associates of one another; **(ii).** If $p \equiv 3 \pmod 4$, then (26.1) has no solution as -1 is a quadratic non-residue of such a prime p by Theorem 6.10; **(iii).** If $p \equiv 1 \pmod 4$, then we already made an investigation in Theorem 26.9. ■

26.5 Quadratic field $\mathbb{Q}(\sqrt{-3})$

In this section, we put $\rho = \frac{1+\sqrt{-3}}{2}$. Recall that $\mathscr{O}_{\mathbb{Q}(\sqrt{-3})} = \mathbb{Z}[\rho]$.

Fact 26.11 $\mathbb{Z}[\rho]$ is a unique factorization domain, i.e. every nonzero nonunit element in $\mathbb{Z}[\rho]$ has a unique (up to reordering and associates) representation as a finite product of irreducible (or equivalently, prime) elements in $\mathbb{Z}[\rho]$.

Let us characterize all irreducible (or equivalently, prime) elements in $\mathbb{Z}[\rho]$. Again, it suffices to consider nonunit factors of rational primes p in $\mathbb{Z}[\rho]$. Note that $N_{\mathbb{Q}(\sqrt{-3})/\mathbb{Q}}(p) = p^2$. Also, if a nonunit $\alpha = a + b\rho \in \mathbb{Z}[\rho]$ with $a, b \in \mathbb{Z}$ is such that $\alpha \mid p$ and that α is not an associate of p, then $N_{\mathbb{Q}(\sqrt{-3})/\mathbb{Q}}(\alpha) = p$, and hence,

$$p = a^2 + ab + b^2, \tag{26.3}$$

or equivalently,

$$4p = (2a+b)^2 + 3b^2. \tag{26.4}$$

For a given rational prime p, if such an α does not exist, i.e. (26.3) or (26.4) has no integer solution (a, b), then we have irreducible elements in $\mathbb{Z}[\rho]$ given by p and its associates. If such an α exists, we assume that $\beta \in \mathbb{Z}[\rho]$ is such that $p = \alpha\beta$, and in this case we have irreducible elements in $\mathbb{Z}[\rho]$ given by α, β and their associates, with duplicates removed.

Theorem 26.12 Let $p \equiv 1 \pmod 6$ be a rational prime. Then the Diophantine equation

$$p = m^2 + mn + n^2 \tag{26.5}$$

has exactly 12 solutions for $m, n \in \mathbb{Z}$.

Proof. By Theorem 7.7, $\left(\frac{-3}{p}\right) = 1$ for $p \equiv 1 \pmod 6$, and hence there is an integer $x \in \mathbb{Z}$ such that $p \mid (x^2+3)$. Further, in $\mathbb{Z}[\rho]$, we have the factorization $x^2 + 3 = (x + \sqrt{-3})(x - \sqrt{-3})$, thereby implying that p is not prime (and hence not irreducible) in $\mathbb{Z}[\rho]$. Thus, we may uniquely factor $p = \alpha\beta$ with $\alpha, \beta \in \mathbb{Z}[\rho]$ nonunit. In particular, if $\alpha = a + b\rho$, then it is plain that $|a| \neq |b|$, that a, b are nonzero, and that neither of α, β are purely real or purely imaginary.

Recall from Theorem 26.4 that the units of $\mathbb{Z}[\rho]$ are ± 1 and $\frac{\pm 1 \pm \sqrt{-3}}{2}$. We claim that β is not an associate of α. If not, then $\alpha^2 = up$ where u is a unit of $\mathbb{Z}[\rho]$. Further, since α is not purely real or purely imaginary, α^2 is not real, and hence $u = \frac{\varepsilon_1}{2} + \frac{\varepsilon_2}{2}\sqrt{-3}$ with $\varepsilon_1, \varepsilon_2 \in \{\pm 1\}$. Note that

$$\alpha^2 = (a + b\rho)^2 = \left((a + \tfrac{b}{2}) + \tfrac{b}{2}\sqrt{-3}\right)^2 = \left(a^2 + ab - \tfrac{b^2}{2}\right) + \left(ab + \tfrac{b^2}{2}\right)\sqrt{-3}.$$

Thus,

$$\begin{cases} a^2 + ab - \frac{b^2}{2} = \frac{\varepsilon_1 p}{2} \\ ab + \frac{b^2}{2} = \frac{\varepsilon_2 p}{2} \end{cases} \iff \begin{cases} a^2 - b^2 = \frac{(\varepsilon_1 - \varepsilon_2)p}{2} \\ a^2 + 2ab = \frac{(\varepsilon_1 + \varepsilon_2)p}{2} \end{cases}.$$

Since $|a| \neq |b|$ and hence $a^2 - b^2 \neq 0$, we only have two possibilities $(\varepsilon_1, \varepsilon_2) = (1, -1)$ or $(-1, 1)$. Thus, $a^2 + 2ab = 0$. Since $a \neq 0$, we get $a + 2b = 0$. Therefore, $\pm p = \frac{(\varepsilon_1 - \varepsilon_2)p}{2} = a^2 - b^2 = 3b^2$. But this is impossible.

We conclude that (26.5) has exactly 12 solutions, determined by the 6 associates of α and the 6 associates of β. ■

Theorem 26.13 The irreducible (or equivalently, prime) elements in $\mathbb{Z}[\rho]$ are:

(i) $\sqrt{-3}$ and its associates;

(ii) rational primes p and their associates for $p = 2$ or $p \equiv 5 \pmod 6$;

(iii) nonunit and nonassociate factors $a + b\rho$ of rational primes p for $p \equiv 1 \pmod 6$, i.e. $a, b \in \mathbb{Z}$ are such that $p = a^2 + ab + b^2$.

Proof. There are four cases: **(i).** If $p = 3$, then we have the factorization $3 = (\sqrt{-3})(-\sqrt{-3})$ where $\sqrt{-3}$ and $-\sqrt{-3}$ are associates of one another; **(ii-a).** If $p = 2$, then it is plain that (26.4) has no solution; **(ii-b).** If $p \equiv 5 \pmod 6$, then (26.4) has no solution as -3 is a quadratic non-residue of such a prime p by Theorem 7.7; **(iii).** If $p \equiv 1 \pmod 6$, then we already made an investigation in Theorem 26.12. ■

27. Continued fractions

27.1 Continued fractions and convergents

Definition 27.1 Let $\{a_0, a_1, a_2, \ldots, a_N\}$ be a finite sequence of numbers. We describe

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots + \cfrac{1}{a_N}}}$$

as a *finite continued fraction.* If the sequence $\{a_0, a_1, a_2, \ldots\}$ is infinite, we say

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots}}$$

is an *infinite continued fraction.* Finite and infinite continued fractions together are called *continued fractions.* We usually adopt the compact notations:

$$a_0 + \frac{1}{a_1 +}\,\frac{1}{a_2 +} \cdots \frac{1}{+\, a_N} \qquad \text{and} \qquad a_0 + \frac{1}{a_1 +}\,\frac{1}{a_2 +} \cdots$$

or

$$\langle a_0, a_1, a_2, \ldots, a_N \rangle \qquad \text{and} \qquad \langle a_0, a_1, a_2, \ldots \rangle.$$

Further, given a continued fraction $\langle a_0, a_1, a_2, \ldots \rangle$, finite or infinite, we call a_n the *n-th (partial) quotient*, $t_n := \langle a_n, a_{n+1}, \ldots \rangle$ the *n-th complete quotient*, and $c_n := \langle a_0, a_1, \ldots, a_n \rangle$ the n-th *convergent.*

R Throughout, if we say $\langle a_0, a_1, a_2, \ldots \rangle$ is a continued fraction, it could be either finite or infinite unless otherwise specified. In the finite case, we usually assume that this continued fraction terminates at a_N, namely, it is given by $\langle a_0, a_1, a_2, \ldots, a_N \rangle$.

Further, when we refer to the index of an expression (e.g., partial quotient, complete quotient, convergent, etc.) related to a finite continued fraction $\langle a_0, a_1, a_2, \ldots, a_N \rangle$, it is automatically understood that this index takes value at most N.

The following proposition is immediate by definition.

Proposition 27.1 Given a continued fraction $\langle a_0, a_1, \ldots \rangle$ and an index n, we have

$$\langle a_0, a_1, \ldots \rangle = \langle a_0, a_1, \ldots, a_{n-1}, \langle a_n, a_{n+1}, \ldots \rangle \rangle.$$

We start with an expression for convergents to a continued fraction.

Theorem 27.2 Let $\langle a_0, a_1, \ldots \rangle$ be a continued fraction. We put

$$\begin{aligned} p_0 &= a_0, & p_1 &= a_0 a_1 + 1, \\ q_0 &= 1, & q_1 &= a_1. \end{aligned}$$

Further, for $k \geq 2$, we define

$$\begin{aligned} p_k &= a_k p_{k-1} + p_{k-2}, \\ q_k &= a_k q_{k-1} + q_{k-2}. \end{aligned}$$

Then for $n \geq 0$,

$$\langle a_0, a_1, \ldots, a_n \rangle = \frac{p_n}{q_n}. \tag{27.1}$$

Proof. We argue by induction on n. It is plain that the relation is true for $n = 0$ and 1. Now assume that it holds for $0, \ldots, n-1$ with $n \geq 2$. By Proposition 27.1, we have $\langle a_0, a_1, \ldots, a_{n-2}, a_{n-1}, a_n \rangle = \langle a_0, a_1, \ldots, a_{n-2}, \langle a_{n-1}, a_n \rangle \rangle = \langle a_0, a_1, \ldots, a_{n-2}, a_{n-1} + \frac{1}{a_n} \rangle$ where the last expression has n partial quotients. We then deduce from the inductive hypothesis that

$$\begin{aligned} \langle a_0, a_1, \ldots, a_{n-2}, a_{n-1}, a_n \rangle &= \langle a_0, a_1, \ldots, a_{n-2}, a_{n-1} + \tfrac{1}{a_n} \rangle \\ &= \frac{(a_{n-1} + \frac{1}{a_n}) p_{n-2} + p_{n-3}}{(a_{n-1} + \frac{1}{a_n}) q_{n-2} + q_{n-3}} \\ &= \frac{a_n (a_{n-1} p_{n-2} + p_{n-3}) + p_{n-2}}{a_n (a_{n-1} q_{n-2} + q_{n-3}) + q_{n-2}} \\ &= \frac{a_n p_{n-1} + p_{n-2}}{a_n q_{n-1} + q_{n-2}} \\ &= \frac{p_n}{q_n}, \end{aligned}$$

as required. ■

Notation 27.2 Given a continued fraction $\langle a_0, a_1, \ldots \rangle$, whenever we say its n-th convergent is $\frac{p_n}{q_n}$, the numbers p_n and q_n are as in Theorem 27.2.

We collect some properties of the numbers p_n and q_n.

Theorem 27.3 For $n \geq 1$, we have

$$p_n q_{n-1} - p_{n-1} q_n = (-1)^{n-1}, \tag{27.2}$$

and equivalently,

$$\frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}} = \frac{(-1)^{n-1}}{q_n q_{n-1}}. \tag{27.3}$$

Proof. This relation simply follows by induction on n. Note that it is trivial for $n = 1$. Further, we find that for $n \geq 2$,

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

thereby yielding the desired result. ■

Theorem 27.4 For $n \geq 2$, we have

$$p_n q_{n-2} - p_{n-2} q_n = (-1)^n a_n, \tag{27.4}$$

and equivalently,

$$\frac{p_n}{q_n} - \frac{p_{n-2}}{q_{n-2}} = \frac{(-1)^n a_n}{q_n q_{n-2}}. \tag{27.5}$$

Proof. We have

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

where (27.2) is applied. ■

Now we shall consider continued fractions with all partial quotients (except a_0) positive.

Theorem 27.5 Let $\langle a_0, a_1, \ldots \rangle$ be a continued fraction where $a_n > 0$ for $n \geq 1$ while a_0 is an arbitrary real number. Then

(i) The even-indexed convergents c_{2k} are strictly monotonically increasing, and the odd-indexed convergents c_{2k+1} are strictly monotonically decreasing.

(ii) Every odd-indexed convergent is greater than any even-indexed convergent.

Proof. It is a trivial observation that $q_n > 0$ for all $n \geq 0$ since $a_n > 0$ for $n \geq 1$.

(i). By (27.5), the sign of $c_n - c_{n-2}$ is determined by $(-1)^n$, and hence the first part follows.

(ii). By (27.3), the sign of $c_n - c_{n-1}$ is determined by $(-1)^{n-1}$. It follows that

$$c_n \begin{cases} < c_{n-1} & \text{if } n \text{ is even,} \\ > c_{n-1} & \text{if } n \text{ is odd.} \end{cases} \tag{27.6}$$

Assume that there exist an even-indexed convergent c_{2i} and an odd-indexed convergent c_{2j+1} such that $c_{2i} \geq c_{2j+1}$. If $2i > 2j+1$ so that $2j+1 \leq 2i-1$, then $c_{2j+1} \geq c_{2i-1}$ since the sequence $\{c_{2k+1}\}$ strictly decreases. Hence, $c_{2i} \geq c_{2i-1}$, but this violates (27.6). On the other hand, if $2i < 2j+1$ so that $2i \leq 2j$, then $c_{2i} \leq c_{2j}$ since the sequence $\{c_{2k}\}$ strictly increases. We then get $c_{2j+1} \leq c_{2j}$, thereby arriving at a contradiction to (27.6) as well. ■

27.2 Simple continued fractions

Definition 27.3 We say $\langle a_0, a_1, \ldots\rangle$ is a *simple continued fraction* if its partial quotients $a_1, a_2, \ldots$ are **positive integers** while a_0 is an **arbitrary integer**. If the sequence of partial quotients is finite, then the simple continued fraction is called *finite*; otherwise, it is called *infinite*. Note that if $\frac{p_n}{q_n}$ is the n-th convergent, then p_n and q_n are integers, and in particular, q_n is positive.

We begin with some quantitative properties for the convergents to a simple continued fraction.

Theorem 27.6 Let $\langle a_0, a_1, \ldots\rangle$ be a simple continued fraction with $\frac{p_n}{q_n}$ the n-th convergent. We have $q_n \geq q_{n-1}$ for $n \geq 1$, where the inequality is strict for $n \geq 2$. In particular, $q_n \geq n$ for $n \geq 0$, where the inequality is strict for $n = 0$ and $n \geq 4$.

Proof. First, we note that $q_0 = 1$ and $q_1 = a_1 \geq 1$. Further, for $n \geq 2$,

$$q_n = a_n q_{n-1} + q_{n-2} \geq q_{n-1} + q_{n-2}.$$

The claims follow as immediate consequences of the above as q_n's are positive integers. ■

Our next object is the *value* of simple continued fractions.

Theorem 27.7 Every finite simple continued fraction $\langle a_0, a_1, \ldots, a_N\rangle$ equals a rational number x, which is called the **value** of $\langle a_0, a_1, \ldots, a_N\rangle$.

Further, the value of $\langle a_0, a_1, \ldots, a_N\rangle$ is greater than any of its even-indexed convergents and less than any of its odd-indexed convergents, except that it equals the last convergent.

Proof. Let $c_n = \frac{p_n}{q_n}$ be the n-th convergent. For the first part, we note that $x = c_N = \frac{p_N}{q_N}$. Since p_N and q_N are integers, we conclude that x is rational.

For the second part, we know from Theorem 27.5(i) that c_N is the maximum among the sequence $\{c_{2k}\}$ if N is even, and the minimum among the sequence $\{c_{2k+1}\}$ if N is odd. The desired statement is a direct consequence of Theorem 27.5(ii). ■

Recall that a rational number $x = \frac{u}{v}$ with u, v integers and $v > 0$ is irreducible if $(u, v) = 1$.

Lemma 27.8 Every convergent to a simple continued fraction is an irreducible rational number.

Proof. Let $\frac{p_n}{q_n}$ be the n-th convergent, and write $d = (p_n, q_n)$. Recalling (27.2), we have $d \mid (p_n q_{n-1} - p_{n-1} q_n) = (-1)^{n-1}$. Thus, $d = 1$, thereby implying that $\frac{p_n}{q_n}$ is irreducible. ■

Regarding infinite simple continued fractions, an important issue that has not yet been considered is their well-definedness.

Theorem 27.9 Every infinite simple continued fraction $\langle a_0, a_1, \ldots\rangle$ converges to an irrational number ξ, which is called the **value** of $\langle a_0, a_1, \ldots\rangle$.

Further, the value of $\langle a_0, a_1, \ldots\rangle$ is greater than any of its even-indexed convergents and less than any of its odd-indexed convergents.

Proof. Let c_n be the n-th convergent. We shall see from Theorem 27.5 that $\{c_{2k}\}_{k\geq 0}$ is a strictly increasing sequence with an upper bound $c_{2k} < c_1$, and that $\{c_{2k+1}\}_{k\geq 0}$ is a strictly decreasing sequence with a lower bound $c_{2k+1} > c_0$. Hence, the two sequences are convergent by the monotone convergence theorem. Further, for each $k \geq 0$, we have

$$c_{2k} < \limsup_{k\to\infty} c_{2k} = \lim_{k\to\infty} c_{2k} =: \xi_e$$

and

$$c_{2k+1} > \liminf_{k\to\infty} c_{2k+1} = \lim_{k\to\infty} c_{2k+1} =: \xi_o.$$

Now we recall Theorems 27.3 and 27.6, and obtain that

$$\left|\frac{p_{2k+1}}{q_{2k+1}} - \frac{p_{2k}}{q_{2k}}\right| = \frac{1}{q_{2k+1}q_{2k}} \leq \frac{1}{2k(2k+1)} \to 0$$

as $k \to \infty$, that is, the sequence $\{c_{2k+1} - c_{2k}\}_{k\geq 0}$ converges to 0. Thus,

$$\xi_o = \lim_{k\to\infty} c_{2k+1} = \lim_{k\to\infty}(c_{2k+1} - c_{2k}) + \lim_{k\to\infty} c_{2k} = 0 + \lim_{k\to\infty} c_{2k} = \xi_e,$$

thereby yielding the convergence of $\{c_k\}_{k\geq 0}$ with

$$\lim_{k\to\infty} c_k = \xi_e = \xi_o =: \xi.$$

By the above arguments, we also have $c_{2k} < \xi_e = \xi$ and $c_{2k+1} > \xi_o = \xi$ for all $k \geq 0$. Finally, the irrationality of ξ will be shown in Corollary 27.17. ■

We close this section by bounding the value of simple continued fractions.

Theorem 27.10 Let $X = \langle a_0, a_1, \ldots\rangle$ be a simple continued fraction. Then $a_0 \leq X \leq a_0 + 1$, where the equality $X = a_0$ occurs if and only if the continued fraction is of the form $\langle a_0\rangle$, and the equality $X = a_0 + 1$ occurs if and only if the continued fraction is of the form $\langle a_0, 1\rangle$. Consequently, X equals an integer if and only if the continued fraction is either $\langle a_0\rangle$ or $\langle a_0, 1\rangle$.

Proof. We start with the finite case $x = \langle a_0, a_1, \ldots, a_N\rangle$, and we shall use induction on N. If $N = 0$, then $x = \langle a_0\rangle = a_0$. If $N = 1$, then $x = a_0 + \frac{1}{a_1}$, and thus $a_0 < x \leq a_0 + 1$ as a_1 is a positive integer. Also, $x = a_0 + 1$ if and only if $a_1 = 1$. For $N \geq 2$, note that

$$x = \langle a_0, a_1, \ldots, a_N\rangle = a_0 + \frac{1}{\langle a_1, \ldots, a_N\rangle}.$$

Let us write $x' = \langle a_1, \ldots, a_N\rangle$, which contains at least two partial quotients. Then $x' > a_1 \geq 1$ by the inductive hypothesis. It turns out that $a_0 < x = a_0 + \frac{1}{x'} < a_0 + 1$.

For infinite simple continued fractions $\xi = \langle a_0, a_1, \ldots\rangle$, we take advantage of the convergents $c_0 = \langle a_0\rangle$ and $c_1 = \langle a_0, a_1\rangle$. By Theorem 27.9, we have $c_0 < \xi < c_1$. Also, the finite case above tells us that $a_0 \leq c_0, c_1 \leq a_0 + 1$. Thus, $a_0 < \xi < a_0 + 1$ follows. ■

27.3 Simple continued fractions of the same value

We begin with finite simple continued fractions.

Definition 27.4 Let $a_0, \ldots, a_N$ be integers with $a_1, \ldots, a_N > 0$. We say

(i) the simple continued fractions $\langle a_0\rangle$ and $\langle a_0 - 1, 1\rangle$ are *companions* of one another;

(ii) the simple continued fractions $\langle a_0, \ldots, a_{N-1}, a_N\rangle$ and $\langle a_0, \ldots, a_{N-1}, a_N - 1, 1\rangle$ are *companions* of one another whenever $N \geq 1$ and $a_N \geq 2$.

Note that every finite simple continued fraction is in exactly one of the four forms.

Proposition 27.11 If two finite simple continued fractions are companions of one another, then they have the same value.

Proof. There are two cases: **(i).** If the two simple continued fractions are $\langle a_0\rangle$ and $\langle a_0-1,1\rangle$, then it is plain that their values are both a_0. **(ii).** If the two simple continued fractions are $\langle a_0,\ldots,a_{N-1},a_N\rangle$ and $\langle a_0,\ldots,a_{N-1},a_N-1,1\rangle$ with $N\geq 1$ and $a_N\geq 2$, then we note that $\langle a_N-1,1\rangle=(a_N-1)+\frac{1}{1}=a_N$, and hence that

$$\langle a_0,\ldots,a_{N-1},a_N-1,1\rangle=\langle a_0,\ldots,a_{N-1},\langle a_N-1,1\rangle\rangle=\langle a_0,\ldots,a_{N-1},a_N\rangle,$$

as required. ■

Definition 27.5 Two finite simple continued fractions $\langle a_0,a_1,\ldots,a_N\rangle$ and $\langle b_0,b_1,\ldots,b_M\rangle$ are called *identical* if $N=M$ and $a_i=b_i$ for all $0\leq i\leq N$.

Theorem 27.12 If two finite simple continued fractions have the same value, then they are either identical or companions of one another.

Proof. Suppose that the two finite simple continued fractions are written as $\langle a_0,a_1,\ldots,a_N\rangle$ and $\langle b_0,b_1,\ldots,b_M\rangle$, and that they are equal to the same value x. Without loss of generality, we assume that $N\leq M$. Let us apply induction on N.

If $N=0$, then $x=\langle a_0\rangle=a_0$ is an integer. Note that in this case, finite simple continued fractions of value equal to x are exactly $\langle x\rangle$ and $\langle x-1,1\rangle$ by Theorem 27.10. It is plain to get the desired claim.

Assume that the claim is true for $0,\ldots,N-1$ with $N\geq 1$. We shall prove the claim for N. In particular, we may assume that x is not an integer for the case where x is an integer was already considered above. Now, again by Theorem 27.10, we have $a_0<x<a_0+1$ and $b_0<x<b_0+1$, indicating that $a_0=b_0$. Noting that

$$\begin{aligned}x=\langle a_0,a_1,\ldots,a_N\rangle&=a_0+\frac{1}{\langle a_1,\ldots,a_N\rangle}\\ =\langle b_0,b_1,\ldots,b_M\rangle&=b_0+\frac{1}{\langle b_1,\ldots,b_M\rangle},\end{aligned}$$

we have $\langle a_1,\ldots,a_N\rangle=\langle b_1,\ldots,b_M\rangle$. By the inductive hypothesis, we know that $\langle a_1,\ldots,a_N\rangle$ and $\langle b_1,\ldots,b_M\rangle$ are either identical or companions of one another, thereby implying the required statement. ■

For infinite simple continued fractions, the consideration is even simpler.

Definition 27.6 Two infinite simple continued fractions $\langle a_0,a_1,\ldots\rangle$ and $\langle b_0,b_1,\ldots\rangle$ are called *identical* if $a_i=b_i$ for all $i\geq 0$.

Theorem 27.13 If two infinite simple continued fractions have the same value, then they are identical.

Proof. The proof is similar to the second half of that for the finite case in Theorem 27.12. Suppose that the two infinite simple continued fractions are written as $\langle a_0,a_1,\ldots\rangle$ and $\langle b_0,b_1,\ldots\rangle$, and that they are equal to the same value ξ. By Theorem 27.10, we have $a_0<\xi<a_0+1$ and $b_0<\xi<b_0+1$, indicating that $a_0=b_0$. Further,

$$\xi=a_0+\frac{1}{\langle a_1,a_2,\ldots\rangle}=b_0+\frac{1}{\langle b_1,b_2,\ldots\rangle}.$$

Hence, $\langle a_1, a_2, \ldots \rangle$ and $\langle b_1, b_2, \ldots \rangle$ have the same value. Repeating the same process gives $a_i = b_i$ for all $i \geq 0$. ■

27.4 Distance from a simple continued fraction to its convergents

It remains unproved in Theorem 27.9 that the value of any infinite simple continued fraction is irrational. For this purpose, a crucial step is to bound the difference between a simple continued fraction and its convergents. We first establish the following relation.

Theorem 27.14 Let $\langle a_0, a_1, \ldots \rangle$ be a continued fraction with $\frac{p_n}{q_n}$ the n-th convergent and t_n the n-th complete quotient. Then for $n \geq 2$,

$$\langle a_0, a_1, \ldots \rangle = \frac{t_n p_{n-1} + p_{n-2}}{t_n q_{n-1} + q_{n-2}}. \tag{27.7}$$

Proof. Note from Proposition 27.1 that $\langle a_0, a_1, \ldots \rangle = \langle a_0, a_1, \ldots, a_{n-1}, t_n \rangle$. For the latter continued fraction, we assume that $\frac{p'_k}{q'_k}$ is its k-th convergent. It is immediate that $p_k = p'_k$ and $q_k = q'_k$ for $0 \leq k \leq n-1$. Now, by Theorem 27.2,

$$\langle a_0, a_1, \ldots, a_{n-1}, t_n \rangle = \frac{p'_n}{q'_n} = \frac{t_n p'_{n-1} + p'_{n-2}}{t_n q'_{n-1} + q'_{n-2}} = \frac{t_n p_{n-1} + p_{n-2}}{t_n q_{n-1} + q_{n-2}},$$

which is exactly the expected relation. ■

Throughout, let $X = \langle a_0, a_1, \ldots \rangle$ be a simple continued fraction with $c_n = \frac{p_n}{q_n}$ the n-th convergent and t_n the n-th complete quotient. In addition, if the simple continued fraction is finite, we write it as $\langle a_0, a_1, \ldots, a_N \rangle$, and add an extra assumption that $N \geq 2$.

Theorem 27.15 For $n \geq 1$ (or $1 \leq n \leq N-1$ in the finite case),

$$X - \frac{p_n}{q_n} = \frac{(-1)^n}{q_n(t_{n+1}q_n + q_{n-1})}. \tag{27.8}$$

Proof. Recalling Theorem 27.14 with n replaced by $n+1$, we have

$$X - \frac{p_n}{q_n} = \frac{t_{n+1}p_n + p_{n-1}}{t_{n+1}q_n + q_{n-1}} - \frac{p_n}{q_n} = \frac{p_{n-1}q_n - p_n q_{n-1}}{q_n(t_{n+1}q_n + q_{n-1})} = \frac{(-1)^n}{q_n(t_{n+1}q_n + q_{n-1})},$$

where we further make use of (27.2) in the last equality. ■

Theorem 27.16 For $n \geq 1$,

$$\frac{1}{q_{n+2}} < |p_n - q_n X| < \frac{1}{q_{n+1}}. \tag{27.9}$$

In the finite case, the same inequalities hold for $1 \leq n \leq N-2$ with the only exception that if $n = N-2$ and $a_N = 1$, then $\frac{1}{q_N} = |p_{N-2} - q_{N-2}X| < \frac{1}{q_{N-1}}$.

Proof. We first deduce from (27.8) that

$$|p_n - q_n X| = \frac{1}{t_{n+1}q_n + q_{n-1}}.$$

Therefore, we need to bound $t_{n+1}q_n + q_{n-1}$. Note that $t_{n+1} = \langle a_{n+1}, a_{n+2}, \ldots \rangle$.

If $n = N-2$ and $a_N = 1$, then $t_{N-1} = \langle a_{N-1}, 1\rangle = a_{N-1} + 1$. Thus,

$$\begin{aligned} t_{N-1}q_{N-2} + q_{N-3} &= (a_{N-1}+1)q_{N-2} + q_{N-3} \\ &= (a_{N-1}q_{N-2} + q_{N-3}) + q_{N-2} \\ &= q_{N-1} + q_{N-2} \\ &= a_N q_{N-1} + q_{N-2} \\ &= q_N. \end{aligned}$$

For $|p_{N-2} - q_{N-2}X| = \frac{1}{q_N} < \frac{1}{q_{N-1}}$, we further recall Theorem 27.6 since $N \geq 2$.

Now we consider the remaining cases. By Theorem 27.10, we always have $a_{n+1} < t_{n+1} < a_{n+1} + 1$. Hence,

$$\begin{aligned} t_{n+1}q_n + q_{n-1} &> a_{n+1}q_n + q_{n-1} \\ &= q_{n+1}. \end{aligned}$$

Also,

$$\begin{aligned} t_{n+1}q_n + q_{n-1} &< (a_{n+1}+1)q_n + q_{n-1} \\ &= q_{n+1} + q_n \\ &\leq a_{n+2}q_{n+1} + q_n \\ &= q_{n+2}. \end{aligned}$$

The desired inequalities therefore follow. ■

Corollary 27.17 The value of any infinite simple continued fraction is an irrational number.

Proof. Let $\langle a_0, a_1, \ldots\rangle$ be an infinite simple continued fraction of value ξ, and assume that its convergents are $\frac{p_n}{q_n}$. We shall prove that ξ is irrational by contradiction. If ξ is rational, we may find integers u and v with $v > 0$ such that $\xi = \frac{u}{v}$. It follows from (27.9) that for all $n \geq 1$,

$$\frac{v}{q_{n+2}} < |vp_n - uq_n| < \frac{v}{q_{n+1}}.$$

Further, by Theorem 27.6, we know that q_n is a positive integer with $q_n \geq n$ for every $n \geq 0$. In the above, if we take $n = v$, then

$$0 < |vp_v - uq_v| < 1.$$

However, $|vp_v - uq_v|$ is an integer, and therefore we arrive at a contradiction. ■

28. Representing real numbers by a simple continued fraction

28.1 Representing rational numbers

From Theorem 27.7, it is known that the value of each finite simple continued fraction equals a rational number. We are also interested in the opposite direction, namely, given an arbitrary rational number, can we express it as a finite simple continued fraction?

Theorem 28.1 Every rational number can be represented as a unique finite simple continued fraction, up to companion.

Proof. The uniqueness has been shown by Theorem 27.12. It suffices to confirm the existence of such a representation as a finite simple continued fraction. Let $x = \frac{u}{v}$ be a rational number where u, v are integers and $v > 0$. Recall the Euclidean Algorithm in which we first put $r_{-1} = u$ and $r_0 = v$:

$$\begin{aligned}
r_{-1} &= a_0 r_0 + r_1, & 0 &< r_1 < r_0;\\
r_0 &= a_1 r_1 + r_2, & 0 &< r_2 < r_1;\\
&\cdots\\
r_{N-2} &= a_{N-1} r_{N-1} + r_N, & 0 &< r_N < r_{N-1};\\
r_{N-1} &= a_N r_N + 0.
\end{aligned}$$

It is immediate that $a_0, \ldots, a_N$ are integers with $a_1, \ldots, a_N > 0$. We shall prove that

$$x = \frac{u}{v} = \langle a_0, a_1, \ldots, a_N \rangle. \tag{28.1}$$

Let $t_n = \langle a_n, a_{n+1}, \ldots, a_N \rangle$ be the n-th complete quotient of $\langle a_0, a_1, \ldots, a_N \rangle$. Our object is to show that for $0 \le n \le N$,

$$t_n = \frac{r_{n-1}}{r_n}. \tag{28.2}$$

Note first that $t_N = \langle a_N \rangle = a_N = \frac{r_{N-1}}{r_N}$, confirming (28.2) for $n = N$. Suppose that (28.2) holds for $n+1, \ldots, N$ with $n \le N-1$. Now we have

$$t_n = a_n + \frac{1}{\langle a_{n+1}, \ldots, a_N \rangle} = a_n + \frac{1}{t_{n+1}} = a_n + \frac{r_{n+1}}{r_n} = \frac{a_n r_n + r_{n+1}}{r_n} = \frac{r_{n-1}}{r_n},$$

as required.

Finally, we take $n=0$ in (28.2), and find that

$$\langle a_0, a_1, \ldots, a_N \rangle = t_0 = \frac{r_{-1}}{r_0} = \frac{u}{v} = x,$$

thereby establishing (28.1). ■

■ **Example 28.1** Consider the finite simple continued fraction representation of $\frac{1071}{462}$. First, the Euclidean Algorithm gives us

$$\begin{aligned} 1071 &= \mathbf{2} \times 462 + 147; \\ 462 &= \mathbf{3} \times 147 + 21; \\ 147 &= \mathbf{7} \times 21 + 0. \end{aligned}$$

It follows that

$$\frac{1071}{462} = \langle 2, 3, 7 \rangle = 2 + \cfrac{1}{3 + \cfrac{1}{7}} = \frac{51}{22}.$$

Also, the companion of $\langle 2,3,7 \rangle$ is $\langle 2,3,6,1 \rangle$, which also has value $\frac{51}{22} = \frac{1071}{462}$. ■

28.2 Representing irrational numbers

A similar treatment can be applied to the representation of irrational numbers by an infinite simple continued fraction, as suggested by Theorem 27.9.

Theorem 28.2 Every irrational number can be represented as a unique infinite simple continued fraction.

Proof. The uniqueness has been shown by Theorem 27.13, and here we only need to examine the existence. Our strategy is a variant of the Euclidean Algorithm used in the rational case, and it is usually called the *Continued Fraction Algorithm.* In particular, we note that each step except the last one in the Euclidean Algorithm can be reformulated as

$$\frac{r_{n-2}}{r_{n-1}} = a_{n-1} + \frac{r_n}{r_{n-1}}, \qquad 0 < \frac{r_n}{r_{n-1}} < 1.$$

Let ξ be an irrational number. We may iteratively compute, with $\xi_0 = \xi$, that

$$\begin{aligned} \xi_0 &= a_0 + \frac{1}{\xi_1}, & \xi_1 &> 1; \\ \xi_1 &= a_1 + \frac{1}{\xi_2}, & \xi_2 &> 1; \\ &\cdots \\ \xi_n &= a_n + \frac{1}{\xi_{n+1}}, & \xi_{n+1} &> 1; \\ &\cdots \end{aligned}$$

Here $a_0, a_1, a_2 \ldots$ are integers, and moreover, $a_1, a_2 \ldots$ are positive. It should be immediately clear that for each $n \geq 0$, we have $a_n = \lfloor \xi_n \rfloor$. Yet another observation is that this procedure will never terminate. This is because if it ends with $\xi_N = a_N + 0$, then ξ_N is rational. Pulling back, we find that ξ_{N-1}, ..., and eventually ξ_0 are rational, thereby getting a contradiction. Now we shall prove that

$$\xi = \langle a_0, a_1, \ldots \rangle. \tag{28.3}$$

We start by showing that for $n \geq 0$,

$$\xi = \langle a_0, a_1, \ldots, a_n, \xi_{n+1} \rangle. \tag{28.4}$$

It is clear that the above relation holds when $n = 0$. Assuming that it is true for $n-1$ with $n \geq 1$, then

$$\xi = \langle a_0, a_1, \ldots, a_{n-1}, \xi_n \rangle = \langle a_0, a_1, \ldots, a_{n-1}, \langle a_n, \xi_{n+1} \rangle \rangle = \langle a_0, a_1, \ldots, a_n, \xi_{n+1} \rangle,$$

thereby establishing (28.4) for n. Letting $\frac{p_k}{q_k}$ be the k-th convergent to $\langle a_0, a_1, \ldots \rangle$, then it is also the k-th convergent to $\langle a_0, a_1, \ldots, a_n, \xi_{n+1} \rangle$ whenever $0 \leq k \leq n$. By Theorem 27.2, we have

$$\xi = \langle a_0, a_1, \ldots, a_n, \xi_{n+1} \rangle = \frac{\xi_{n+1} p_n + p_{n-1}}{\xi_{n+1} q_n + q_{n-1}}.$$

Now, by (27.2),

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

Recalling Theorem 27.6 and noting that $\xi_{n+1} > 1$, we further get

$$\left| \xi - \frac{p_n}{q_n} \right| = \frac{1}{q_n (\xi_{n+1} q_n + q_{n-1})} < \frac{1}{n^2} \to 0,$$

as $n \to \infty$. It follows that

$$\xi = \lim_{n \to \infty} \frac{p_n}{q_n} = \langle a_0, a_1, \ldots \rangle.$$

Thus, (28.3) is established and the existence of an infinite simple continued fraction of value ξ is confirmed. ■

■ **Example 28.2** Consider the infinite simple continued fraction representation of $\sqrt{3}$. First, the Continued Fraction Algorithm gives us

$$\begin{aligned}
\sqrt{3} &= \mathbf{1} + \frac{1}{\frac{1}{2}(1+\sqrt{3})}; \\
\frac{1}{2}(1+\sqrt{3}) &= \mathbf{1} + \frac{1}{1+\sqrt{3}}; \\
1+\sqrt{3} &= \mathbf{2} + \frac{1}{\frac{1}{2}(1+\sqrt{3})}; \\
\frac{1}{2}(1+\sqrt{3}) &= \mathbf{1} + \frac{1}{1+\sqrt{3}}; \\
1+\sqrt{3} &= \mathbf{2} + \frac{1}{\frac{1}{2}(1+\sqrt{3})}; \\
&\cdots
\end{aligned}$$

It follows that

$$\sqrt{3} = \langle 1, 1, 2, 1, 2, \ldots \rangle = \langle 1, \overline{1, 2} \rangle.$$

Note that in this infinite simple continued fraction, there is a repeating portion $\overline{1, 2}$. ■

28.3 Periodic simple continued fractions

As we have seen in Example 28.2, the infinite simple continued fraction representation of $\sqrt{3}$ contains a repeating portion. Now we shall give a systematic study of such continued fractions.

Definition 28.1 A *periodic simple continued fraction* is a simple continued fraction that eventually repeats. More precisely, if $\langle a_0, a_1, \dots \rangle$ is periodic, then there exist an index r and a positive integer m such that $a_i = a_j$ whenever $i, j \geq r$ and $i \equiv j \pmod{m}$. Furthermore, if

$$\langle a_0, a_1, \dots, a_{r-1}, a_r, \dots, a_{r+m-1}, a_r, \dots, a_{r+m-1}, \dots \rangle$$

is a periodic simple continued fraction with $a_r, \dots, a_{r+m-1}$ repeating, we write it as

$$\langle a_0, a_1, \dots, a_{r-1}, \overline{a_r, \dots, a_{r+m-1}} \rangle.$$

R Note that given a periodic simple continued fraction, there are different ways to express it. For instance, for $\langle 1,2,3,2,3,2,3\dots \rangle$ with $2,3$ repeating, we can write it as

$$\langle 1,2,3,2,3,2,3\dots \rangle = \langle 1, \overline{2,3} \rangle = \langle 1, \overline{2,3,2,3} \rangle = \langle 1,2,\overline{3,2} \rangle = \langle 1,2,3,\overline{2,3} \rangle = \cdots.$$

Definition 28.2 Given a periodic simple continued fraction, its *period* is the smallest possible number of partial quotients in a repeating portion. So in the above instance, the period is 2.

Theorem 28.3 Let $\langle a_0, a_1, \dots \rangle$ be a periodic simple continued fraction of period m_0. If it is expressed as

$$\langle a_0, a_1, \dots, a_{r_1-1}, \overline{a_{r_1}, \dots, a_{r_1+m_1-1}} \rangle,$$

then we have $m_0 \mid m_1$.

Proof. Since $\langle a_0, a_1, \dots \rangle$ has period m_0, we may also express it as

$$\langle a_0, a_1, \dots, a_{r_0-1}, \overline{a_{r_0}, \dots, a_{r_0+m_0-1}} \rangle.$$

Let us put $r = \max\{r_0, r_1\}$ and $m = (m_0, m_1)$. We shall show that in $\langle a_0, a_1, \dots \rangle$, there is a repeating portion that contains m partial quotients. Now since m_0 is the smallest length of any choice of repeating portions, we should have $m_0 \leq m = (m_0, m_1)$, thereby implying that $m_0 \mid m_1$.

To prove the above claim, we shall show that for $i \geq r$ and $k \geq 0$, it is always true that $a_i = a_{i+km}$. Noting that $m = (m_0, m_1)$, by Theorem 2.5 there exist integers u and v such that $m = um_0 + vm_1$. In particular, at least one of u and v is positive. If u is positive, then noting that $i + km \geq r \geq r_1$, that $i + kum_0 \geq r \geq r_1$ and that $i + km = i + k(um_0 + vm_1) \equiv i + kum_0 \pmod{m_1}$, we have $a_{i+km} = a_{i+kum_0}$. Further, we have $i \geq r \geq r_0$ and $i + kum_0 \geq r \geq r_0$, while $i + kum_0 \equiv i \pmod{m_0}$. Hence, $a_i = a_{i+kum_0} = a_{i+km}$, as required. For the case where v is positive, we proceed with a similar analysis. ■

We say an irrational number is *quadratic* if it is a root of a quadratic polynomial with rational coefficients. In other words, its minimal polynomial over $\mathbb{Q}$ is of degree 2, and thus itself is of degree 2 over $\mathbb{Q}$. Based on the discussions in Sect. 26.1, we know that such a number can be expressed as $u + v\sqrt{d}$ where $d > 1$ is squarefree and u, v are rational numbers.

Theorem 28.4 The value of any periodic simple continued fraction is a quadratic irrational number.

Proof. Let

$$\xi = \langle a_0, a_1, \dots, a_{r-1}, \overline{b_0, \dots, b_{m-1}} \rangle$$

be a periodic simple continued fraction, and put

$$\eta = \langle \overline{b_0, \dots, b_{m-1}} \rangle.$$

Note that

$$\xi = \langle a_0, a_1, \dots, a_{r-1}, \eta \rangle.$$

As long as we can show that $\eta \in \mathbb{Q}(\sqrt{d})$ for a certain squarefree $d > 1$, it is immediate that $\xi \in \mathbb{Q}(\sqrt{d})$, thereby establishing the desired result.

It is known that η is irrational by Theorem 27.9. Now we show that η is of degree 2 over $\mathbb{Q}$, and hence confirm the claim. To see this, we observe that

$$\begin{aligned} \eta = \langle \overline{b_0, \dots, b_{m-1}} \rangle &= \langle b_0, \dots, b_{m-1}, b_0, \dots, b_{m-1}, \overline{b_0, \dots, b_{m-1}} \rangle \\ &= \langle b_0, \dots, b_{m-1}, b_0, \dots, b_{m-1}, \eta \rangle. \qquad (28.5) \end{aligned}$$

Here we tacitly repeat the portion $b_0, \dots, b_{m-1}$ twice in the last continued fraction in (28.5) as we want to make sure that there are at least two partial quotients before the last η. Let $\frac{p_n}{q_n}$ be the n-th convergent to $\langle \overline{b_0, \dots, b_{m-1}} \rangle$. Then

$$\eta = \frac{\eta p_{2m-1} + p_{2m-2}}{\eta q_{2m-1} + q_{2m-2}}.$$

Thus, η is a root of the following quadratic polynomial over $\mathbb{Q}$:

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

Since η is irrational, and hence is not of degree 1 over $\mathbb{Q}$, we arrived at the desired claim. ■

28.4 Representing quadratic irrational numbers

One may again ask if the infinite simple continued fraction representation of a quadratic irrational number is periodic or not. The object of the current section is to answer this question.

Theorem 28.5 Every quadratic irrational number can be represented as a unique periodic simple continued fraction.

Proof. Note that by Theorem 28.2, the infinite simple continued fraction representation of any quadratic irrational number is unique. Hence it is sufficient to show that this continued fraction is periodic.

Let $\xi = u + v\sqrt{d}$ be a quadratic irrational number where $d > 1$ is squarefree and $u, v \in \mathbb{Q}$. Assume that it satisfies

$$c_2 x^2 + c_1 x + c_0 = 0. \qquad (28.6)$$

The above quadratic equation has two irrational roots: $u + v\sqrt{d}$ and $u - v\sqrt{d}$.

Suppose that the infinite simple continued fraction representation of ξ is given by

$$\xi = \langle a_0, a_1, \dots \rangle,$$

with $\frac{p_n}{q_n}$ the n-th convergent and t_n the n-th complete quotient.

Throughout, let $n \geq 3$. By (27.7),

$$\xi = \frac{t_n p_{n-1} + p_{n-2}}{t_n q_{n-1} + q_{n-2}}.$$

Substituting the above into (28.6) gives

$$A_n t_n^2 + B_n t_n + C_n = 0, \tag{28.7}$$

where

$$\begin{aligned}
A_n &= c_2 p_{n-1}^2 + c_1 p_{n-1} q_{n-1} + c_0 q_{n-1}^2,\\
B_n &= 2c_2 p_{n-1} p_{n-2} + c_1 p_{n-1} q_{n-2} + c_1 p_{n-2} q_{n-1} + 2c_0 q_{n-1} q_{n-2},\\
C_n &= c_2 p_{n-2}^2 + c_1 p_{n-2} q_{n-2} + c_0 q_{n-2}^2.
\end{aligned}$$

In particular,

$$A_n = C_{n+1}.$$

Note that $A_n \neq 0$ for $n \geq 2$ since otherwise

$$c_2 p_{n-1}^2 + c_1 p_{n-1} q_{n-1} + c_0 q_{n-1}^2 = 0,$$

then

$$c_2 \left(\frac{p_{n-1}}{q_{n-1}}\right)^2 + c_1 \left(\frac{p_{n-1}}{q_{n-1}}\right) + c_0 = 0,$$

thereby contradicting the fact that the solutions to (28.6) are irrational.

Now our goal is to bound $|A_n|$, $|B_n|$ and $|C_n|$ by constants independent of n. First, we deduce from (27.9) that for $k \geq 1$,

$$\left|\xi - \frac{p_k}{q_k}\right| < \frac{1}{q_k q_{k+1}} < \frac{1}{q_k^2}.$$

Hence, we shall write

$$\xi - \frac{p_k}{q_k} = -\frac{\varepsilon_k}{q_k^2},$$

where $|\varepsilon_k| < 1$ depends on k. It turns out that

$$\begin{aligned}
C_n &= q_{n-2}^2 \left(c_2 \left(\xi + \frac{\varepsilon_{n-2}}{q_{n-2}^2}\right)^2 + c_1 \left(\xi + \frac{\varepsilon_{n-2}}{q_{n-2}^2}\right) + c_0 \right)\\
&= q_{n-2}^2 (c_2 \xi^2 + c_1 \xi + c_0) + 2c_2 \varepsilon_{n-2} \xi + \frac{c_2 \varepsilon_{n-2}^2}{q_{n-2}^2} + c_1 \varepsilon_{n-2}\\
&= 2c_2 \varepsilon_{n-2} \xi + \frac{c_2 \varepsilon_{n-2}^2}{q_{n-2}^2} + c_1 \varepsilon_{n-2}.
\end{aligned}$$

Thus,

$$|C_n| \leq 2|c_2 \xi| + |c_2| + |c_1|.$$

Recalling that $A_n = C_{n+1}$, we also have

$$|A_n| \leq 2|c_2 \xi| + |c_2| + |c_1|.$$

Finally, we compute that

$$\begin{aligned}
B_n^2 - 4A_n C_n &= \left(c_1^2 - 4c_0 c_2\right)\left(p_{n-1} q_{n-2} - p_{n-2} q_{n-1}\right)^2\\
&= c_1^2 - 4c_0 c_2,
\end{aligned}$$

where we make use of (27.2). It follows that

$$\begin{aligned} B_n^2 &\leq |c_1^2 - 4c_0c_2| + 4|A_nC_n| \\ &\leq |c_1^2 - 4c_0c_2| + 4\big(2|c_2\xi| + |c_2| + |c_1|\big)^2. \end{aligned}$$

Since $|A_n|$, $|B_n|$ and $|C_n|$ are uniformly bounded, it follows from the pigeonhole principle that there must exist three distinct $n_1, n_2, n_3 \geq 3$ such that

$$(A_{n_1}, B_{n_1}, C_{n_1}) = (A_{n_2}, B_{n_2}, C_{n_2}) = (A_{n_3}, B_{n_3}, C_{n_3})$$

and that at least two of t_{n_1}, t_{n_2} and t_{n_3} are equal. Assuming that $t_{n_1} = t_{n_2}$ with $n_1 < n_2$, that is,

$$\langle a_{n_1}, a_{n_1+1}, \ldots\rangle = \langle a_{n_2}, a_{n_2+1}, \ldots\rangle,$$

then by Theorem 27.13, we have $a_{n_1+k} = a_{n_2+k}$ for all $k \geq 0$. It turns out that

$$\xi = \langle a_0, a_1, \ldots, a_{n_1-1}, \overline{a_{n_1}, \ldots, a_{n_2-1}}\rangle,$$

which is periodic, as required. ■

28.5 Purely periodic simple continued fractions

We close this lecture by investigating a special type of periodic simple continued fractions.

Definition 28.3 A periodic simple continued fraction is called *purely periodic* if the initial non-repeating block is not present. That is, the continued fraction is of the form

$$\langle\overline{a_0, a_1, \ldots, a_{m-1}}\rangle.$$

The following result is due to Évariste Galois (*Ann. Math. Pures Appl. [Ann. Gergonne]* **19** (1828/29), 294–301).

Theorem 28.6 (Galois). Let $\xi = u + v\sqrt{d}$ be a quadratic irrational number where $d > 1$ is squarefree and $u, v \in \mathbb{Q}$ and put $\xi' = u - v\sqrt{d}$, the conjugate of ξ over $\mathbb{Q}$. Then ξ is represented by a purely periodic simple continued fraction if and only if $\xi > 1$ and $-1 < \xi' < 0$.

Proof. We start with necessity. Assume that the purely periodic simple continued fraction representation of ξ is given by

$$\xi = \langle\overline{a_0, a_1, \ldots, a_{m-1}}\rangle.$$

Then $a_0 \geq 1$ as it appears again in the continued fraction. By Theorem 27.10, we have $\xi > a_0 \geq 1$. Letting $\frac{p_n}{q_n}$ be the n-th convergent to $\langle\overline{a_0, a_1, \ldots, a_{m-1}}\rangle$, then from the proof of Theorem 28.4, we see that ξ is a root of

$$f(x) = q_{2m-1}x^2 + (q_{2m-2} - p_{2m-1})x - p_{2m-2} \in \mathbb{Z}[x].$$

Since ξ' is conjugate to ξ over $\mathbb{Q}$, we know that it is also a root of $f(x)$. Now it suffices to show that $f(x)$ has a root in the interval $(-1, 0)$. To see this, we simply note that

$$f(-1) = (p_{2m-1} - p_{2m-2}) + (q_{2m-1} - q_{2m-2}) > 0,$$

and that

$$f(0) = -p_{2m-2} < 0.$$

For sufficiency, we argue by contradiction. Suppose on the contrary that the periodic simple continued fraction representation of ξ is not purely periodic and write it as

$$\begin{aligned}\xi &= \langle a_0, a_1, \ldots\rangle \\ &= \langle a_0, a_1, \ldots, a_{r-1}, \overline{a_r, \ldots, a_{r+m-1}}\rangle,\end{aligned}$$

where $r \geq 1$ and $a_{r-1} \neq a_{(r-1)+m}$. Letting $\eta = \langle\overline{a_r, \ldots, a_{r+m-1}}\rangle$, then $\xi = \langle a_0, a_1, \ldots, a_{r-1}, \eta\rangle$. It follows that $\eta \in \mathbb{Q}(\sqrt{d})$ since $\xi \in \mathbb{Q}(\sqrt{d})$. Further, if t_n denotes the n-th complete quotient of $\langle a_0, a_1, \ldots, a_{r-1}, \overline{a_r, \ldots, a_{r+m-1}}\rangle$, then there is an index $\ell \geq n$ such that

$$t_n = \langle a_n, \ldots, a_\ell, \overline{a_r, \ldots, a_{r+m-1}}\rangle = \langle a_n, \ldots, a_\ell, \eta\rangle,$$

and hence $t_n \in \mathbb{Q}(\sqrt{d})$. Let us write $t_n = u_n + v_n\sqrt{d}$ with $u_n, v_n \in \mathbb{Q}$ and put $t'_n = u_n - v_n\sqrt{d}$, the conjugate of t_n over $\mathbb{Q}$. Since $t_n = \langle a_n, a_{n+1}, \ldots\rangle = \langle a_n, t_{n+1}\rangle$, we have

$$t_n = a_n + \frac{1}{t_{n+1}}.$$

Taking conjugates over $\mathbb{Q}$ on both sides of the above gives

$$t'_n = a_n + \frac{1}{t'_{n+1}}$$

so that

$$t'_{n+1} = \frac{1}{t'_n - a_n}.$$

Noting that $t_0 = \xi$ and hence that $t'_0 = \xi'$, we have $-1 < t'_0 < 0$. Since $a_n \geq 1$ for all $n \geq 0$ as $\xi > 1$, we inductively get $-1 < t'_n < 0$ for all n. Finally, since

$$t_r = t_{r+m} = \langle\overline{a_r, \ldots, a_{r+m-1}}\rangle,$$

we have

$$\begin{aligned}t_{r-1} - t_{(r-1)+m} &= \left(a_{r-1} + \frac{1}{t_r}\right) - \left(a_{(r-1)+m} + \frac{1}{t_{r+m}}\right) \\ &= a_{r-1} - a_{(r-1)+m}.\end{aligned}$$

Hence, $t_{r-1} - t_{(r-1)+m}$ is a nonzero integer, and so is its conjugate $t'_{r-1} - t'_{(r-1)+m}$ over $\mathbb{Q}$. However, we have shown that $-1 < t'_{r-1}, t'_{(r-1)+m} < 0$, thereby indicating that $-1 < t'_{r-1} - t'_{(r-1)+m} < 1$. This leads to a contradiction. ■

■ **Example 28.3** **(i).** Consider the simple continued fraction representation of $3+\sqrt{13}$. Note that $3+\sqrt{13} > 1$ and $-1 < 3-\sqrt{13} < 0$. The Continued Fraction Algorithm gives us that

$$3+\sqrt{13} = \langle\overline{6,1,1,1,1}\rangle,$$

which is purely periodic.

(ii). Consider the simple continued fraction representation of $5+\sqrt{22}$. Note that $5+\sqrt{22} > 1$ but $5-\sqrt{22} > 0$. The Continued Fraction Algorithm gives us that

$$5+\sqrt{22} = \langle 9, \overline{1,2,4,2,1,8}\rangle,$$

which is periodic but not purely periodic. ■

29. Approximations of irrational numbers

29.1 Approximation exponents

It is known that rational numbers are dense in real numbers. In other words, given any $x \in \mathbb{R}$, we may always find a rational number $\frac{p}{q}$ with p, q integers and $q > 0$ such that $|x - \frac{p}{q}|$ is arbitrarily small. On the other hand, for rational numbers $\frac{p}{q}$ with a fixed denominator q, it is always possible to find a numerator p such that $|x - \frac{p}{q}| < \frac{1}{q}$. This is because the disjoint intervals $\left[\frac{p}{q}, \frac{p+1}{q}\right)$ $(p \in \mathbb{Z})$ cover $\mathbb{R}$, and x falls into exactly one of these intervals. In the case where x is a rational number, say in the irreducible expression $\frac{a}{b}$ with a, b integers, $b > 0$ and $(a, b) = 1$, if we assume that q is not a multiple of b, then $x \neq \frac{p}{q}$ for any choice of p. The above arguments imply that for any real x, there exist infinitely many integer pairs (p, q) with $q > 0$ such that

$$0 < \left|x - \frac{p}{q}\right| < \frac{1}{q}. \tag{29.1}$$

Usually, if we want to approximate x by a rational number $\frac{p}{q}$, we shall expect that the approximation behaves better than simply satisfying (29.1). To measure how rapid an approximation is, we introduce the concept of *approximation exponent.*

Definition 29.1 Let x be a real number. The *approximation exponent*, also known as the *irrationality exponent*, of x, denoted by $\mu(x)$, is defined to be the supremum of the set of real numbers μ such that the inequalities

$$0 < \left|x - \frac{p}{q}\right| < \frac{1}{q^{\mu}} \tag{29.2}$$

hold for an infinite number of integer pairs (p, q) with $q > 0$.

R Here we exclude the zero-error case, i.e. we require that $|x - \frac{p}{q}| > 0$ as we do not want to approximate a rational number by itself. For instance, if $x = \frac{a}{b}$ is rational and irreducible, we may find infinitely many integer pairs $(p, q) = (ka, kb)$ with k positive integers such that $|x - \frac{p}{q}| = 0 < \frac{1}{q^{\mu}}$ for any real μ. But then all $\frac{p}{q}$ point to the same value $\frac{a}{b} = x$, and such approximations do not make too much sense. An anecdote regarding this issue is due to Arnold Ross, of the Ohio State University, who used to ask, "*What is an approximation to 5?*" and then answer, "*Any number other than 5.*"

Note that the discussions regarding (29.1) can be paraphrased as follows.

Proposition 29.1 Let x be a real number.

(i) We have $\mu(x) \geq 1$.

(ii) For every $\mu \geq 1$, if we fix $q > 0$ an integer, then there are at most two integers p such that (29.2) holds.

It is easy to determine the approximation exponent for rational numbers.

Theorem 29.2 Every rational number has approximation exponent equal to 1.

Proof. Let $\frac{a}{b}$ be a rational number with a, b integers and $b > 0$. Assume that $\varepsilon > 0$ is arbitrary. We want to determine all rational approximations $\frac{p}{q}$ to $\frac{a}{b}$ such that

$$0 < \left|\frac{a}{b} - \frac{p}{q}\right| < \frac{1}{q^{1+\varepsilon}}. \tag{29.3}$$

Note that

$$\left|\frac{a}{b} - \frac{p}{q}\right| = \frac{|aq - bp|}{bq} \geq \frac{1}{bq}.$$

Hence,

$$\frac{1}{q^{1+\varepsilon}} > \frac{1}{bq},$$

that is,

$$q < b^{\frac{1}{\varepsilon}},$$

which is bounded. Further, for each q, there are at most two p such that (29.3) holds. Thus, $\mu(\frac{a}{b}) < 1 + \varepsilon$ for any $\varepsilon > 0$, thereby implying that $\mu(\frac{a}{b}) = 1$. ∎

As we have seen above, the approximation to a rational number is almost trivial and hence less interesting. But what happens when we approximate an irrational number? Is there a criterion for its approximation exponent? Further, if in (29.2), μ is taken to be the approximation exponent, can we determine the corresponding approximations $\frac{p}{q}$? Such questions bring about a now flourishing area in Number Theory known as the *Theory of Diophantine Approximations.*

Throughout, let ξ be irrational unless otherwise specified.

29.2 Approximations by convergents

Assume that ξ has the infinite simple continued fraction representation $\langle a_0, a_1, \ldots \rangle$ with $\frac{p_n}{q_n}$ its n-th convergent. The following result is an immediate implication of Theorem 27.16.

Theorem 29.3 For $n \geq 1$,

$$0 < \left|\xi - \frac{p_n}{q_n}\right| < \frac{1}{q_n^2}. \tag{29.4}$$

Consequently, for every irrational number ξ, we have $\mu(\xi) \geq 2$.

We shall show that the above approximation by convergents is in some sense the *best.*

Theorem 29.4 Let $n \geq 2$. For any rational number $\frac{p}{q} \neq \frac{p_n}{q_n}$ with $0 < q \leq q_n$, we have

$$|p_n - q_n\xi| < |p - q\xi| \tag{29.5}$$

and

$$\left|\xi - \frac{p_n}{q_n}\right| < \left|\xi - \frac{p}{q}\right|. \tag{29.6}$$

Proof. Note that (29.5) implies (29.6). This is because if we assume (29.5), then

$$\left|\xi - \frac{p_n}{q_n}\right| < \frac{|p - q\xi|}{q_n} \leq \frac{|p - q\xi|}{q} = \left|\xi - \frac{p}{q}\right|.$$

Now we prove (29.5). Our starting point is Theorem 27.16, which tells us that

$$|p_n - q_n\xi| < \frac{1}{q_{n+1}} < |p_{n-1} - q_{n-1}\xi|.$$

Hence, it suffices to show (29.5) for $q_{n-1} < q \leq q_n$. We further assume that $(p,q) = 1$ as if $(p,q) = d > 1$, we have $|p - q\xi| = d|p' - q'\xi| > |p' - q'\xi|$ where $p = dp'$ and $q = dq'$.

If $q = q_n$, then since $\frac{p}{q} \neq \frac{p_n}{q_n}$, we have $p \neq p_n$. Hence,

$$|p_n - p| \geq 1.$$

On the other hand, we know from Theorems 27.6 and 27.16 that

$$|p_n - q_n\xi| < \frac{1}{q_{n+1}} < \frac{1}{2}.$$

It follows that

$$|p - q\xi| = |p - q_n\xi| \geq |p_n - p| - |p_n - q_n\xi| > \frac{1}{2} > |p_n - q_n\xi|.$$

Now suppose that $q_{n-1} < q < q_n$. Recall from Lemma 27.8 that both $\frac{p_n}{q_n}$ and $\frac{p_{n-1}}{q_{n-1}}$ are in the irreducible expression. Also, $\frac{p}{q}$ is irreducible as we have assumed that $(p,q) = 1$. By the uniqueness of irreducible expressions in Theorem 19.2, we know that the three numbers $\frac{p}{q}$, $\frac{p_n}{q_n}$ and $\frac{p_{n-1}}{q_{n-1}}$ are distinct. Assuming that u and v are such that

$$\begin{cases} p = up_n + vp_{n-1}, \\ q = uq_n + vq_{n-1}, \end{cases}$$

and recalling from (27.2) that $p_nq_{n-1} - p_{n-1}q_n = (-1)^{n-1}$, we solve the above system and get

$$\begin{cases} u = (-1)^{n-1}\big(pq_{n-1} - p_{n-1}q\big), \\ v = (-1)^{n}\big(pq_n - p_nq\big). \end{cases}$$

Thus, u and v are nonzero integers. Since $0 < q_{n-1} < q = uq_n + vq_{n-1} < q_n$, we find that u and v have different signs. Further, by Theorem 27.9, the numbers $p_n - q_n\xi$ and $p_{n-1} - q_{n-1}\xi$ have different signs. So $u \cdot (p_n - q_n\xi)$ and $v \cdot (p_{n-1} - q_{n-1}\xi)$ have the same sign. Now we have

$$p - q\xi = (up_n + vp_{n-1}) - (uq_n + vq_{n-1})\xi = u \cdot (p_n - q_n\xi) + v \cdot (p_{n-1} - q_{n-1}\xi),$$

so that

$$|p - q\xi| = |u \cdot (p_n - q_n\xi)| + |v \cdot (p_{n-1} - q_{n-1}\xi)| > |p_n - q_n\xi|,$$

as required. ■

We are also able to further elaborate the upper bound in (29.4).

Theorem 29.5 For $n \geq 0$, at least one of $\frac{p_n}{q_n}$ and $\frac{p_{n+1}}{q_{n+1}}$ is such that

$$0 < \left|\xi - \frac{p}{q}\right| < \frac{1}{2q^2}. \tag{29.7}$$

Proof. We start by noting from (27.2) that

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

Supposing on the contrary that

$$\left|\xi - \frac{p_n}{q_n}\right| \geq \frac{1}{2q_n^2} \qquad \text{and} \qquad \left|\xi - \frac{p_{n+1}}{q_{n+1}}\right| \geq \frac{1}{2q_{n+1}^2},$$

and noting that $\xi - \frac{p_n}{q_n}$ and $\xi - \frac{p_{n+1}}{q_{n+1}}$ have different signs, we have

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

Thus,

$$(q_n - q_{n+1})^2 \leq 0.$$

The only possibility is $n = 0$ and $q_0 = q_1 = 1$, from which we get $a_1 = 1$. However, in this case we still have

$$\begin{aligned}
0 < \left|\xi - \frac{p_1}{q_1}\right| &= \left|\langle a_0, 1, a_2, a_3, \ldots\rangle - \frac{a_0+1}{1}\right| \\
&= \left|a_0 + \frac{1}{1 + \frac{1}{\langle a_2, a_3, \ldots\rangle}} - (a_0+1)\right| \\
&= 1 - \frac{1}{1 + \frac{1}{\langle a_2, a_3, \ldots\rangle}} \\
&< \frac{1}{2} = \frac{1}{2q_1^2},
\end{aligned}$$

where we use the fact that $\langle a_2, a_3, \ldots\rangle > 1$. ■

Finally, we show that the bounds in (29.7) indeed provide a characterization of convergents.

Theorem 29.6 If an irreducible rational number $\frac{p}{q}$ is such that

$$0 < \left|\xi - \frac{p}{q}\right| < \frac{1}{2q^2}, \tag{29.8}$$

then $\frac{p}{q}$ is a convergent to the simple continued fraction representation of ξ.

Proof. Write

$$\xi - \frac{p}{q} = \frac{(-1)^m\theta}{q^2}, \qquad (0 < \theta < \tfrac{1}{2}).$$

Recall from Theorem 28.1 that $\frac{p}{q}$ has exactly two finite simple continued fraction representations, and that they are companions of one another so that the numbers of partial quotients in the two continued fractions differ by 1. We may choose one representation

$$\frac{p}{q} = \langle a_0, a_1, \ldots, a_N \rangle$$

such that $(-1)^N = (-1)^m$.

Let η be a real number such that

$$\xi = \langle a_0, a_1, \ldots, a_N, \eta \rangle.$$

If we can show that $\eta > 1$, then in the simple continued fraction representation of η, say $\eta = \langle b_0, b_1, \ldots \rangle$, we must have $b_0 > 0$ by the Continued Fraction Algorithm as $b_0 = \lfloor \eta \rfloor$. Thus, ξ is represented by the simple continued fraction

$$\xi = \langle a_0, a_1, \ldots, a_N, \langle b_0, b_1, \ldots \rangle \rangle = \langle a_0, a_1, \ldots, a_N, b_0, b_1, \ldots \rangle,$$

which is also the unique representation by Theorem 28.2. Further, $\frac{p}{q}$ is the convergent $\frac{p_N}{q_N}$.

Now we prove that $\eta > 1$. If $N = 0$ so that $(-1)^m = (-1)^N = 1$, then $\xi = \frac{p}{q} + \frac{\theta}{q^2}$. Also, $\frac{p}{q} = \langle a_0 \rangle = a_0$ is an integer. Thus, $\eta = \frac{q^2}{\theta} > 1$, as required. Suppose that $N \geq 1$ in what follows. By Theorem 27.2,

$$\xi = \frac{\eta p_N + p_{N-1}}{\eta q_N + q_{N-1}},$$

so that

$$\eta = -\frac{\xi q_{N-1} - p_{N-1}}{\xi q_N - p_N}.$$

Recalling that $\xi = \frac{p}{q} + \frac{(-1)^m \theta}{q^2} = \frac{p_N}{q_N} + \frac{(-1)^N \theta}{q_N^2}$, we have

$$\begin{aligned}
\eta &= -\frac{(-1)^N \theta q_{N-1} + q_N(p_N q_{N-1} - p_{N-1} q_N)}{(-1)^N \theta q_N} \\
&= -\frac{(-1)^N \theta q_{N-1} + q_N(-1)^{N-1}}{(-1)^N \theta q_N} \\
&= \frac{q_N - \theta q_{N-1}}{\theta q_N} \\
&= \frac{1}{\theta} - \frac{q_{N-1}}{q_N} \\
&> 1,
\end{aligned}$$

where we use the facts that $0 < \theta < \frac{1}{2}$ and that $q_N \geq q_{N-1} > 0$ in the last inequality. ■

29.3 Dirichlet's approximation theorem

Let us repeat the statement in Theorem 29.3.

Theorem 29.7 For any irrational number ξ, there exist infinitely many rational numbers $\frac{p}{q}$ such that

$$0 < \left| \xi - \frac{p}{q} \right| < \frac{1}{q^2}. \tag{29.9}$$

This result is originally due to Dirichlet, who provided an ingenious proof based on the following approximation theorem.

Theorem 29.8 (Dirichlet's Approximation Theorem). For any real number x and any integer $Q \geq 1$, there exist integers p and q with $1 \leq q \leq Q$ such that

$$\left| x - \frac{p}{q} \right| \leq \frac{1}{q(Q+1)}. \tag{29.10}$$

Consequently, by casting out all common factors of p and q, the above inequality also holds with the additional condition that $(p,q) = 1$.

Proof. We start by dividing the interval $[0,1)$ into $Q+1$ disjoint intervals

$$I_k := \left[\frac{k-1}{Q+1}, \frac{k}{Q+1} \right), \qquad (1 \leq k \leq Q+1).$$

Consider the Q numbers

$$\{x\}, \{2x\}, \ldots, \{Qx\},$$

where $\{x\} = x - \lfloor x \rfloor$. It is known that each number falls into exactly one of the above intervals. There are three cases:

(i). Suppose that there exists an integer q with $1 \leq q \leq Q$ such that $\{qx\} \in I_1$, i.e. $0 \leq qx - \lfloor qx \rfloor < \frac{1}{Q+1}$. Then we take $p = \lfloor qx \rfloor$ and find that (29.10) holds.

(ii). Suppose that there exists an integer q with $1 \leq q \leq Q$ such that $\{qx\} \in I_{Q+1}$, i.e. $\frac{Q}{Q+1} \leq qx - \lfloor qx \rfloor < 1$. Then we take $p = \lfloor qx \rfloor + 1$ and find that (29.10) holds.

(iii). Suppose that none of the integers q with $1 \leq q \leq Q$ are such that $\{qx\} \in I_1 \cup I_{Q+1}$ so the Q numbers distribute over the remaining $Q-1$ intervals. Then by the pigeonhole principle, there exist two distinct integers q_1 and q_2 with $1 \leq q_1 < q_2 \leq Q$ such that $\{q_1 x\}$ and $\{q_2 x\}$ fall into the same interval. Hence, $|\{q_2 x\} - \{q_1 x\}| < \frac{1}{Q+1}$. Noting that $\{q_2 x\} - \{q_1 x\} = (q_2 - q_1)x - (\lfloor q_2 x \rfloor - \lfloor q_1 x \rfloor)$, we choose $q = q_2 - q_1$ and $p = \lfloor q_2 x \rfloor - \lfloor q_1 x \rfloor$, and find that (29.10) holds. ■

Now we are in a position to reproduce Dirichlet's proof of Theorem 29.7.

Dirichlet's Proof of Theorem 29.7. Since ξ is irrational, we know that $|\xi - \frac{p}{q}| > 0$ for any rational number $\frac{p}{q}$. It is also clear from Theorem 29.8 that there exists at least one $\frac{p}{q}$ such that (29.9) holds, for the reason that we may choose an arbitrary $Q \geq 1$ together with a rational $\frac{p}{q}$ with $1 \leq q \leq Q$ such that $|\xi - \frac{p}{q}| \leq \frac{1}{q(Q+1)} < \frac{1}{q^2}$.

Now assuming that we are given a list of distinct rational numbers $\{\frac{p_1}{q_1}, \ldots, \frac{p_n}{q_n}\}$ where each number satisfies (29.9), our object is to construct a new rational number $\frac{p_{n+1}}{q_{n+1}}$ different from any in the previous list so that (29.9) is still valid, thereby yielding the infinitude of such rational numbers. To do so, we choose an integer Q^* where

$$Q^* > \max \left\{ \left| \xi - \frac{p_1}{q_1} \right|^{-1}, \ldots, \left| \xi - \frac{p_n}{q_n} \right|^{-1} \right\}.$$

By Theorem 29.8, we can find a rational number $\frac{p_{n+1}}{q_{n+1}}$ with $1 \leq q_{n+1} \leq Q^*$ such that

$$\left| \xi - \frac{p_{n+1}}{q_{n+1}} \right| \leq \frac{1}{q_{n+1}(Q^*+1)} < \frac{1}{q_{n+1}^2}.$$

It remains to show that $\frac{p_{n+1}}{q_{n+1}}$ is distinct from any number in $\{\frac{p_1}{q_1}, \ldots, \frac{p_n}{q_n}\}$. But this is trivial since for any $1 \leq k \leq n$, we have

$$\left| \xi - \frac{p_{n+1}}{q_{n+1}} \right| \leq \frac{1}{q_{n+1}(Q^*+1)} < \frac{1}{Q^*} < \left| \xi - \frac{p_k}{q_k} \right|.$$

Therefore, $\frac{p_{n+1}}{q_{n+1}} \neq \frac{p_k}{q_k}$ for any $1 \leq k \leq n$. ■

29.4 Liouville's approximation theorem

Dirichlet's approximation theorem presents us with a lower bound for the approximation exponent of irrational numbers. For the opposite direction, we have an important result due to the French mathematician Joseph Liouville which bounds from above the approximation exponent of algebraic numbers.

Recall that an algebraic number is a root of a polynomial with integer coefficients, and its degree is the degree of its minimal polynomial.

Theorem 29.9 (Liouville's Approximation Theorem). Let ξ be an algebraic number of degree n. Then there is a positive constant c, depending only on ξ, such that

$$\left|\xi - \frac{p}{q}\right| > \frac{c}{q^n}, \tag{29.11}$$

whenever p,q are integers, $q > 0$ and $\frac{p}{q} \neq \xi$, while if $n \geq 2$, the last condition can be omitted. Consequently, we have $\mu(\xi) \leq n$.

Proof. Let $f(x) = a_n x^n + \cdots + a_1 x + a_0 \in \mathbb{Z}[x]$ be such that $f(\xi) = 0$. There are two cases.

(i). If $|\xi - \frac{p}{q}| > 1$, we may choose $c_1 = 1$ so that

$$\left|\xi - \frac{p}{q}\right| > 1 \geq \frac{1}{q^n} = \frac{c_1}{q^n}.$$

(ii). If $0 < |\xi - \frac{p}{q}| \leq 1$, we start with the observation that $q^n f(\frac{p}{q})$ is an integer. It is also plain that $f(\frac{p}{q}) \neq 0$. Suppose on the contrary that $f(\frac{p}{q}) = 0$. Then $(x - \frac{p}{q})$ divides $f(x)$ in $\mathbb{Q}[x]$, and hence we get a polynomial of degree $n-1$, given by

$$\frac{f(x)}{x - p/q} \in \mathbb{Q}[x],$$

such that ξ is also its root as we have assumed that $\xi \neq \frac{p}{q}$. But this violates the fact that ξ is of degree n. Thus,

$$q^n \left| f\left(\frac{p}{q}\right)\right| \geq 1.$$

On the other hand, we know from the mean value theorem that there exists a number η between ξ and $\frac{p}{q}$ such that

$$f'(\eta) = \frac{f(\xi) - f(\frac{p}{q})}{\xi - \frac{p}{q}} = -\frac{f(\frac{p}{q})}{\xi - \frac{p}{q}},$$

where we use $f(\xi) = 0$. Since $0 < |\xi - \frac{p}{q}| \leq 1$, we further find that η is in the closed interval $[\xi - 1, \xi + 1]$. Obviously, $f'(x)$ is bounded on this interval for $f'(x)$ is continuous. In particular, we may assume that $|f'(x)| < M$ for all $x \in [\xi - 1, \xi + 1]$ where $M > 0$ depends only on ξ. Hence, $|f'(\eta)| < M$. It follows that

$$\left|\xi - \frac{p}{q}\right| > \frac{\left|f(\frac{p}{q})\right|}{M} \geq \frac{1}{Mq^n} = \frac{c_2}{q^n},$$

where we put $c_2 = \frac{1}{M}$.

Combining the two cases, we may choose the constant c as $c = \min\{c_1, c_2\}$, which depends only on ξ, such that (29.11) holds. ■

Corollary 29.10 Every quadratic irrational number has approximation exponent equal to 2.

Proof. Dirichlet's approximation theorem asserts that the approximation exponent of any irrational number is at least 2, while from Liouville's approximation theorem, the approximation exponent of any quadratic algebraic number is at most 2. The claim therefore follows. ■

Can we say more about the approximation exponent of an arbitrary algebraic number? Along this direction one shall encounter a highlight in the Theory of Diophantine Approximations, which is due to the British mathematician Klaus Roth (*Mathematika* **2** (1955), 1–20). For this contribution, Roth was awarded the Fields Medal in 1958.

Roth's Theorem Every algebraic number that is not rational has approximation exponent equal to 2.

29.5 Transcendence revisited

One important application of Liouville's approximation theorem concerns the construction of transcendental numbers. Here we give an instance.

Theorem 29.11 The number

$$\kappa = \sum_{m\geq 0} \frac{1}{2^{m!}}$$

is transcendental.

Proof. Suppose on the contrary that κ is an algebraic number of degree n. By Liouville's approximation theorem, there is a constant c such that for all rational numbers $\frac{p}{q} \neq \kappa$,

$$\left|\kappa - \frac{p}{q}\right| > \frac{c}{q^n}.$$

Now consider the numbers $\frac{p(M)}{q(M)}$ for M positive integers where

$$p(M) = \sum_{m=0}^{M} 2^{M!-m!} \qquad \text{and} \qquad q(M) = 2^{M!}.$$

We find that

$$0 < \kappa - \frac{p(M)}{q(M)} = \sum_{m\geq M+1} \frac{1}{2^{m!}} < \frac{1}{2^{(M+1)!}} \sum_{\ell\geq 0} \frac{1}{2^{\ell}} = \frac{2}{q(M)^{M+1}}.$$

To ensure the assumption made from Liouville's approximation theorem, we must have

$$\frac{2}{q(M)^{M+1}} > \frac{c}{q(M)^n},$$

that is,

$$2^{(M+1-n)\cdot M!} < \frac{2}{c}.$$

However, the left-hand side goes to infinity as M goes to infinity, thereby yielding a contradiction. ■

30. Pell's equation

30.1 Pell's equation

Definition 30.1 *Pell's equation* is of the form

$$x^2 - Dy^2 = 1, \tag{30.1}$$

where the integer $D > 1$, which is not a square, is given. Whenever we refer to a solution (x,y) to Pell's equation, we assume that x and y are **integers** unless otherwise specified. That is to say, we are usually only interested in the **integer solutions**.

Pell's equation is named after the English mathematician John Pell, but this attribution, which should belong to the English mathematician William Brouncker as the first European to solve Pell's equation, was mistakenly arisen by Euler. However, this equation was studied even earlier outside Europe. For instance, in the Indian mathematician Brahmagupta's work *Brāhmasphuṭasiddhānta*, an integer solution to $x^2 - 92y^2 = 1$ was discovered.

Definition 30.2 Pell's equation $x^2 - Dy^2 = 1$ always has two solutions $(x,y) = (\pm 1, 0)$. The two solutions are called *trivial solutions*. All other solutions are called *nontrivial solutions*.

Now we start with some basic facts about the nontrivial solutions to Pell's equation.

Fact 30.1 Let (x,y) be a nontrivial solution to Pell's equation $x^2 - Dy^2 = 1$.

(i) x and y are coprime.

(ii) $(x,-y)$, $(-x,y)$, $(-x,-y)$ are also nontrivial solutions to $x^2 - Dy^2 = 1$.

(iii) $|x| \geq 2$ and $|y| \geq 1$.

(iv) $x + y\sqrt{D} \in \begin{cases} (1,+\infty) & \text{if } x > 0 \text{ and } y > 0, \\ (0,1) & \text{if } x > 0 \text{ and } y < 0, \\ (-1,0) & \text{if } x < 0 \text{ and } y > 0, \\ (-\infty,-1) & \text{if } x < 0 \text{ and } y < 0. \end{cases}$

(v) If (x_1,y_1) and (x_2,y_2) with $x_1,y_1,x_2,y_2 > 0$ are two distinct solutions, then the three inequalities $x_1 > x_2$, $y_1 > y_2$ and $x_1 + y_1\sqrt{D} > x_2 + y_2\sqrt{D}$ are equivalent.

Recall that we have assumed that the integer $D > 1$ is not a square. This means that

$\sqrt{D}$ is irrational, and hence that 1 and $\sqrt{D}$ are linearly independent over $\mathbb{Q}$. In other words, if we have $a_1 + b_1\sqrt{D} = a_2 + b_2\sqrt{D}$ with $a_1, b_1, a_2, b_2 \in \mathbb{Q}$, then $a_1 = a_2$ and $b_1 = b_2$. Further, the set $\{a + b\sqrt{D} : a, b \in \mathbb{Q}\}$ forms a field under addition and multiplication, and indeed it is $\mathbb{Q}(\sqrt{D})$. We shall freely use these facts below.

Theorem 30.2 Let (x,y) and (x',y') be solutions to Pell's equation $x^2 - Dy^2 = 1$. Then the integer pair (X,Y) given by

$$X + Y\sqrt{D} = (x + y\sqrt{D})(x' + y'\sqrt{D})$$

gives a solution to the same equation. In particular, for k positive integers, the integer pairs (x_k, y_k) are solutions to the equation where

$$x_k + y_k\sqrt{D} = (x + y\sqrt{D})^k.$$

Consequently, if there is a nontrivial solution, then there are infinitely many solutions.

Proof. From an algebraic perspective, this theorem is trivial as we may invoke the norms

$$\begin{aligned} X^2 - DY^2 &= N_{\mathbb{Q}(\sqrt{D})/\mathbb{Q}}(X + Y\sqrt{D}) \\ &= N_{\mathbb{Q}(\sqrt{D})/\mathbb{Q}}(x + y\sqrt{D}) \cdot N_{\mathbb{Q}(\sqrt{D})/\mathbb{Q}}(x' + y'\sqrt{D}) \\ &= (x^2 - Dy^2)(x'^2 - Dy'^2) \\ &= 1. \end{aligned}$$

However, we may also directly compute that

$$\begin{cases} X = xx' + yy'D, \\ Y = xy' + x'y, \end{cases}$$

so that

$$\begin{aligned} X^2 - DY^2 &= (xx' + yy'D)^2 - D(xy' + x'y)^2 \\ &= (x^2 - Dy^2)(x'^2 - Dy'^2) \\ &= 1. \end{aligned}$$

For the second claim, we simply apply induction on k. Finally, if there is a nontrivial solution (x,y), then we may assume that $x, y > 0$ so that $x + y\sqrt{D} > 1$. Hence $(x + y\sqrt{D})^k$ increases strictly with k, thereby implying that the pairs (x_k, y_k) are distinct. ■

30.2 Existence of solutions

Note that the above discussions are built on the assumption that Pell's equation $x^2 - Dy^2 = 1$ has a nontrivial solution. However, the core question is does such a solution *exist*? To address an answer, we require a lemma based on Dirichlet's approximation theorem.

Lemma 30.3 Let $D > 1$ be an integer that is not a square. Then there exists an integer t with $0 < |t| < 2\sqrt{d}$ such that there are infinitely many integer pairs (p,q) with

$$p^2 - Dq^2 = t. \tag{30.2}$$

Proof. Recall that $\sqrt{D}$ is irrational. By repeatedly applying Theorem 29.8, we may get an infinite sequence of integer triples (p_1,q_1,Q_1), (p_2,q_2,Q_2), $\ldots$ where $1 \le q_n \le Q_n$ such that

$$\left|\sqrt{D}-\frac{p_n}{q_n}\right| \le \frac{1}{q_n(Q_n+1)} \qquad \text{and} \qquad Q_{n+1} > \max\left\{Q_n, \left|\sqrt{D}-\frac{p_n}{q_n}\right|^{-1}\right\}.$$

Note that the numbers Q_n form a strictly increasing sequence. Also, we claim that the pairs (p_n,q_n) are distinct. This is because if $1 \le i < j$ are two different indices, then

$$\left|\sqrt{D}-\frac{p_i}{q_i}\right| > \frac{1}{Q_{i+1}} > \cdots > \frac{1}{Q_j} > \frac{1}{q_j(Q_j+1)} \ge \left|\sqrt{D}-\frac{p_j}{q_j}\right|.$$

Further, we see that for every $n \ge 1$,

$$\begin{aligned}|p_n^2-Dq_n^2| &= |p_n-q_n\sqrt{D}|\cdot|p_n+q_n\sqrt{D}|\\ &= |p_n-q_n\sqrt{D}|\cdot|p_n-q_n\sqrt{D}+2q_n\sqrt{D}|\\ &\le \frac{1}{Q_n+1}\left(\frac{1}{Q_n+1}+2Q_n\sqrt{D}\right)\\ &< 2\sqrt{D}.\end{aligned}$$

By the pigeonhole principle, there is an integer t with $|t| < 2\sqrt{D}$ such that there are infinitely many pairs (p,q) among (p_n,q_n) with $p^2-Dq^2=t$. Further, this t cannot be 0 as D is not a square. ■

Theorem 30.4 Let $D > 1$ be an integer that is not a square. Then Pell's equation

$$x^2-Dy^2=1$$

has infinitely many solutions.

Proof. By Theorem 30.2, it suffices to show the existence of one nontrivial solution. Let t be as in Lemma 30.3 and assume that (P,Q) and (P',Q') with $P,Q,P',Q' > 0$ are two distinct solutions to

$$p^2-Dq^2=t$$

such that

$$P \equiv P' \pmod{|t|} \qquad \text{and} \qquad Q \equiv Q' \pmod{|t|}.$$

Let

$$X+Y\sqrt{D}=\frac{P'+Q'\sqrt{D}}{P+Q\sqrt{D}},$$

so that

$$X+Y\sqrt{D}=\frac{(P'+Q'\sqrt{D})(P-Q\sqrt{D})}{(P+Q\sqrt{D})(P-Q\sqrt{D})}=\frac{(PP'-DQQ')+(PQ'-P'Q)\sqrt{D}}{t},$$

and hence that

$$\begin{cases} X=\dfrac{PP'-DQQ'}{t},\\ Y=\dfrac{PQ'-P'Q}{t}.\end{cases}$$

Note that $PP' - DQQ' \equiv P^2 - DQ^2 = t \equiv 0 \pmod{|t|}$ and that $PQ' - P'Q \equiv PQ - PQ = 0 \pmod{|t|}$. Hence, X and Y are integers. Also, $Y \neq 0$ as if we assume on the contrary that $Y = 0$, then $\frac{P}{Q} = \frac{P'}{Q'}$ so that

$$\frac{t}{Q^2} = \left(\frac{P}{Q}\right)^2 - D = \left(\frac{P'}{Q'}\right)^2 - D = \frac{t}{Q'^2},$$

and hence that $Q = Q'$ and $P = P'$, which contradicts the assumption that (P,Q) and (P',Q') are distinct. Finally,

$$\begin{aligned} X^2 - DY^2 &= N_{\mathbb{Q}(\sqrt{D})/\mathbb{Q}}(X + Y\sqrt{D}) \\ &= \frac{N_{\mathbb{Q}(\sqrt{D})/\mathbb{Q}}(P' + Q'\sqrt{D})}{N_{\mathbb{Q}(\sqrt{D})/\mathbb{Q}}(P + Q\sqrt{D})} \\ &= \frac{P'^2 - DQ'^2}{P^2 - DQ^2} \\ &= 1. \end{aligned}$$

We conclude that (X,Y) is the desired nontrivial solution. ■

30.3 Structure of solutions

As long as we have shown the existence of solutions to Pell's equation, it becomes natural to ask if there is a way to characterize all the solutions. It turns out that we may go further and consider integer solutions (x,y) to a slightly generalized variant of Pell's equation

$$x^2 - Dy^2 = \pm 1,$$

where $D > 1$ is not a square. The solutions $(\pm 1, 0)$ are still called *trivial* while all other solutions are *nontrivial*. Again, we have some basic facts about nontrivial solutions.

Fact 30.5 Let (x,y) be a nontrivial solution to the equation $x^2 - Dy^2 = \pm 1$.

(i) x and y are coprime.
(ii) $(x,-y)$, $(-x,y)$, $(-x,-y)$ are also nontrivial solutions to $x^2 - Dy^2 = \pm 1$.
(iii) If $x + y\sqrt{D} > 1$, then $x > 0$ and $y > 0$.

The structure of the solutions to $x^2 - Dy^2 = \pm 1$ can be characterized as follows.

Theorem 30.6 Let $D > 1$ be an integer that is not a square. Then the equation

$$x^2 - Dy^2 = \pm 1 \tag{30.3}$$

has a nontrivial solution (x_1, y_1) with $x_1, y_1 > 0$ such that $x_1 + y_1\sqrt{D}$ is minimal. Further, all solutions to this equation are given by $(\pm x_k, \pm y_k)$ with $k \geq 0$ where

$$x_k + y_k\sqrt{D} = (x_1 + y_1\sqrt{D})^k, \tag{30.4}$$

so that

$$x_k = \frac{(x_1 + y_1\sqrt{D})^k + (x_1 - y_1\sqrt{D})^k}{2} \quad \text{and} \quad y_k = \frac{(x_1 + y_1\sqrt{D})^k - (x_1 - y_1\sqrt{D})^k}{2\sqrt{D}}. \tag{30.5}$$

In particular, the case where $k = 0$ corresponds to the trivial solutions $(\pm 1, 0)$.

Consequently, the following statements are true.

(i) If the solution (x_1, y_1) is such that $x_1^2 - Dy_1^2 = 1$, then all solutions to Pell's equation $x^2 - Dy^2 = 1$ are given by $(\pm x_k, \pm y_k)$ with $k \geq 0$, and the equation $x^2 - Dy^2 = -1$ has no solution.

(ii) If the solution (x_1, y_1) is such that $x_1^2 - Dy_1^2 = -1$, then all solutions to Pell's equation $x^2 - Dy^2 = 1$ are given by $(\pm x_{2k}, \pm y_{2k})$ with $k \geq 0$, and all solutions to the equation $x^2 - Dy^2 = -1$ are given by $(\pm x_{2k+1}, \pm y_{2k+1})$ with $k \geq 0$.

Definition 30.3 The solution (x_1, y_1) in this theorem is called the *fundamental solution* to (30.3).

Proof. We have shown that $x^2 - Dy^2 = 1$ has a nontrivial solution, say (X, Y) with $X, Y > 0$. Now there are only finitely many solutions (x, y) to $x^2 - Dy^2 = \pm 1$ such that $x + y\sqrt{D}$ lies in the interval $(1, X + Y\sqrt{d}]$. This claim is due to the fact that x is bounded by $0 < x \leq X$ and hence that $0 < y \leq \sqrt{(X^2+1)/D}$; otherwise, if $x > X$, then $Dy^2 = x^2 \mp 1 > X^2 - 1 = DY^2$ so that $y > Y$ and hence that $x + y\sqrt{D} > X + Y\sqrt{D}$. Thus, we can find a solution (x_1, y_1) with $x_1, y_1 > 0$ so that $x_1 + y_1\sqrt{D}$ is minimal. It is plain that all $(\pm x_k, \pm y_k)$ are solutions to the equation by a similar argument to that for Theorem 30.2. Also, (30.5) follows from the fact that $x_k - y_k\sqrt{D} = (x_1 - y_1\sqrt{D})^k$, by taking conjugates over $\mathbb{Q}$.

Now assume that there is a solution $(\tilde{x}, \tilde{y})$ with $\tilde{x}, \tilde{y} > 0$ to $x^2 - Dy^2 = \pm 1$ such that it is not among (x_k, y_k). Then we may find an index $n \geq 1$ such that

$$(x_1 + y_1\sqrt{D})^n = x_n + y_n\sqrt{D} < \tilde{x} + \tilde{y}\sqrt{D} < x_{n+1} + y_{n+1}\sqrt{D} = (x_1 + y_1\sqrt{D})^{n+1}.$$

Note that $0 < |x_1 - y_1\sqrt{D}| < 1$. Multiplying by $|x_1 - y_1\sqrt{D}|^n$ to each part of the above gives

$$1 < (\tilde{x} + \tilde{y}\sqrt{D}) \cdot |x_1 - y_1\sqrt{D}|^n < x_1 + y_1\sqrt{D}.$$

Let us define

$$\tilde{x}_1 + \tilde{y}_1\sqrt{D} = (\tilde{x} + \tilde{y}\sqrt{D}) \cdot |x_1 - y_1\sqrt{D}|^n,$$

so that $\tilde{x}_1$ and $\tilde{y}_1$ are integers. Now,

$$\begin{aligned}\tilde{x}_1^2 - D\tilde{y}_1^2 &= N_{\mathbb{Q}(\sqrt{D})/\mathbb{Q}}(\tilde{x}_1 + \tilde{y}_1\sqrt{D}) \\ &= N_{\mathbb{Q}(\sqrt{D})/\mathbb{Q}}(\tilde{x} + \tilde{y}\sqrt{D}) \cdot N_{\mathbb{Q}(\sqrt{D})/\mathbb{Q}}|x_1 - y_1\sqrt{D}|^n \\ &= \pm 1,\end{aligned}$$

implying that $(\tilde{x}_1, \tilde{y}_1)$ is also a solution. Since $\tilde{x}_1 + \tilde{y}_1\sqrt{D} > 1$, we have $\tilde{x}_1, \tilde{y}_1 > 0$. Then we are led to a contradiction as $\tilde{x}_1 + \tilde{y}_1\sqrt{D} < x_1 + y_1\sqrt{D}$ violates the minimality of $x_1 + y_1\sqrt{D}$. In other words, we cannot have the fabled solution $(\tilde{x}, \tilde{y})$.

For the last conclusion, we simply use the fact that $x_k^2 - Dy_k^2 = (x_1^2 - Dy_1^2)^k$. ■

■ **Example 30.1** **(i).** Consider $D = 2$, i.e. the equation $x^2 - 2y^2 = \pm 1$. A direct computation gives $(x_1, y_1) = (1, 1)$ with $x_1^2 - 2y_1^2 = 1 - 2 = -1$. Hence, we have

$$x_k = \frac{(1+\sqrt{2})^k + (1-\sqrt{2})^k}{2} \quad \text{and} \quad y_k = \frac{(1+\sqrt{2})^k - (1-\sqrt{2})^k}{2\sqrt{2}}. \tag{30.6}$$

The solutions to $x^2 - 2y^2 = 1$ are given by $(\pm x_{2k}, \pm y_{2k})$ with $k \geq 0$, e.g. $(\pm 1, 0)$, $(\pm 3, \pm 2)$, etc. The solutions to $x^2 - 2y^2 = -1$ are given by $(\pm x_{2k+1}, \pm y_{2k+1})$ with $k \geq 0$, e.g. $(\pm 1, \pm 1)$, $(\pm 7, \pm 5)$, etc.

(ii). Consider $D = 3$, i.e. the equation $x^2 - 3y^2 = \pm 1$. A direct computation gives $(x_1, y_1) = (2, 1)$ with $x_1^2 - 3y_1^2 = 4 - 3 = 1$. Hence, we have

$$x_k = \frac{(2+\sqrt{3})^k + (2-\sqrt{3})^k}{2} \quad \text{and} \quad y_k = \frac{(2+\sqrt{3})^k - (2-\sqrt{3})^k}{2\sqrt{3}}. \tag{30.7}$$

The solutions to $x^2 - 3y^2 = 1$ are given by $(\pm x_k, \pm y_k)$ with $k \geq 0$, e.g. $(\pm 1, 0)$, $(\pm 2, \pm 1)$, $(\pm 7, \pm 4)$, etc. The equation $x^2 - 3y^2 = -1$ has no solution. ■

30.4 Units of real quadratic fields

Now we are in a position to complete the proof of the characterization of units of real quadratic fields as claimed in Sect. 26.1.

Theorem 30.7 (Units of Real Quadratic Fields). Let $d > 1$ be a squarefree integer in $\mathbb{Z}$. Then the units of $\mathscr{O}_{\mathbb{Q}(\sqrt{d})}$ are real. Further, there exists a unique unit $\varepsilon > 1$ such that all units of $\mathscr{O}_{\mathbb{Q}(\sqrt{d})}$ are of the form $\pm\varepsilon^n$ with $n \in \mathbb{Z}$.

Proof. According to Lemma 26.5, it is sufficient to study the equation

$$a^2 - db^2 = \pm 1, \tag{30.8}$$

and characterize all solutions (a, b) with $a, b \in \frac{1}{2}\mathbb{Z}$. Since the units are of the form $a + b\sqrt{d}$ by Lemma 26.5, they are real numbers.

The rest has the same logic as that for Theorem 30.6. In particular, enlarging the domain of solutions from $\mathbb{Z}$ to $\frac{1}{2}\mathbb{Z}$ will not affect the finitude of solutions (a, b) to $a^2 - db^2 = \pm 1$ such that $a + b\sqrt{d}$ lies in the interval $(1, A + B\sqrt{d}]$, where (A, B) with $A, B > 0$ is again a fixed solution to Pell's equation $a^2 - db^2 = 1$. Now the desired unit $\varepsilon = a_1 + b_1\sqrt{d}$ is determined by the solution (a_1, b_1) in $\frac{1}{2}\mathbb{Z}$ with $a_1, b_1 > 0$ so that $a_1 + b_1\sqrt{d}$ is minimal and hence unique. ■

30.5 Fundamental solution via continued fractions

Our last object is to find an efficient way to determine the fundamental solution to (30.3). Recall that $\sqrt{D}$ is a quadratic irrational number, so by Theorem 28.5 it has a unique periodic simple continued fraction representation.

Theorem 30.8 Assuming that the periodic simple continued fraction representation of $\sqrt{D}$ has period m, then it is of the form $\langle a_0, \overline{a_1, \ldots, a_m} \rangle$.

Proof. By the Continued Fraction Algorithm, we have $a_0 = \lfloor \sqrt{D} \rfloor$. Let us write $\sqrt{D} = \langle a_0, \delta \rangle$ where δ is a real number. Then $\sqrt{D} = a_0 + \frac{1}{\delta} = \lfloor \sqrt{D} \rfloor + \frac{1}{\delta}$ so that $\delta = \frac{1}{\sqrt{D} - \lfloor \sqrt{D} \rfloor}$. Now the conjugate of δ over $\mathbb{Q}$ is $\delta' = \frac{1}{-\sqrt{D} - \lfloor \sqrt{D} \rfloor}$. Since $\delta > 1$ and $-1 < \delta' < 0$, we conclude from Theorem 28.6 that δ can be represented by a purely periodic simple continued fraction, and hence that $\sqrt{D}$ has the required continued fraction representation. ■

Throughout, we always assume that the periodic simple continued fraction representation of $\sqrt{D}$, which has period m, is written as $\langle a_0, a_1, \ldots \rangle = \langle a_0, \overline{a_1, \ldots, a_m} \rangle$. Let $\frac{p_n}{q_n}$ be its n-th convergent and t_n its n-th complete quotient.

Theorem 30.9 If $(x,y)=(p,q)$ with $p,q>0$ is a solution to $x^2-Dy^2=\pm 1$, then $\frac{p}{q}$ is a convergent to $\sqrt{D}=\langle a_0,a_1,\ldots\rangle$.

Proof. Note that $p^2=Dq^2\pm 1\geq Dq^2-1\geq q^2$ where we use the fact that $D\geq 2$ and $q\geq 1$ in the last inequality. Hence, $p\geq q$. Further, $\pm 1=p^2-Dq^2=(p+q\sqrt{D})(p-q\sqrt{D})$ implies that

$$0<\left|\sqrt{D}-\frac{p}{q}\right|=\frac{1}{q(p+q\sqrt{D})}\leq\frac{1}{q(q+q\sqrt{D})}=\frac{1}{(1+\sqrt{D})q^2}<\frac{1}{2q^2}.$$

Since p and q are coprime, Theorem 29.6 tells us that $\frac{p}{q}$ is a convergent to $\sqrt{D}$. ■

Theorem 30.10 Assume that the periodic simple continued fraction representation of $\sqrt{D}$ has period m. Then $p_n^2-Dq_n^2=\pm 1$ if and only if $m\mid(n+1)$. Further, whenever $m\mid(n+1)$, we have

$$p_n^2-Dq_n^2=(-1)^{n+1}. \tag{30.9}$$

Consequently, (p_{m-1},q_{m-1}) is the fundamental solution to $x^2-Dy^2=\pm 1$, and in particular,

$$p_{m-1}^2-Dq_{m-1}^2=(-1)^m. \tag{30.10}$$

Proof. We start with sufficiency. Noting that $t_{n+2}=t_{m\cdot\frac{n+1}{m}+1}=t_1=\frac{1}{\sqrt{D}-a_0}$, we obtain from Theorem 27.14 that

$$\sqrt{D}=\frac{t_{n+2}p_{n+1}+p_n}{t_{n+2}q_{n+1}+q_n}=\frac{p_{n+1}+p_n(\sqrt{D}-a_0)}{q_{n+1}+q_n(\sqrt{D}-a_0)}.$$

Multiplying both sides by $q_{n+1}+q_n(\sqrt{D}-a_0)$ and treating them as linear combinations of 1 and $\sqrt{D}$, we derive by equating coefficients that

$$\begin{cases}0=p_{n+1}-p_na_0-q_nD,\\ 0=q_{n+1}-q_na_0-p_n.\end{cases}$$

Eliminating a_0 and recalling (27.2), we have

$$p_n^2-Dq_n^2=-(p_{n+1}q_n-p_nq_{n+1})=(-1)^{n+1},$$

as required.

For necessity, it is enough to prove that $t_{n+2}=t_1$, as if this is the case, then we may express $\sqrt{D}=\langle a_0,a_1,\ldots\rangle$ as $\langle a_0,\overline{a_1,\ldots,a_{n+1}}\rangle$, and Theorem 28.3 asserts that $m\mid(n+1)$. Now we prove this claim. Note that the sign of $p_n^2-Dq_n^2$ is determined by $\frac{p_n}{q_n}-\sqrt{D}$ whose sign is $(-1)^{n+1}$ as indicated by Theorem 27.9. Hence, $p_n^2-Dq_n^2=(-1)^{n+1}$. If $n=0$, then $p_0^2-Dq_0^2=-1$. Also, $p_0=a_0=\lfloor\sqrt{D}\rfloor$ and $q_0=1$, so that $\lfloor\sqrt{D}\rfloor^2-D=-1$. Recall that

$$t_1=\frac{1}{\sqrt{D}-\lfloor\sqrt{D}\rfloor}=\frac{\sqrt{D}+\lfloor\sqrt{D}\rfloor}{D-\lfloor\sqrt{D}\rfloor^2}=\sqrt{D}+\lfloor\sqrt{D}\rfloor.$$

Therefore,

$$t_2=\frac{1}{t_1-\lfloor t_1\rfloor}=\frac{1}{(\sqrt{D}+\lfloor\sqrt{D}\rfloor)-\lfloor\sqrt{D}+\lfloor\sqrt{D}\rfloor\rfloor}=\frac{1}{\sqrt{D}-\lfloor\sqrt{D}\rfloor},$$

thereby yielding $t_1 = t_2 = t_{0+2}$, as required. Suppose that $n \geq 1$ in what follows. Since

$$\sqrt{D} = \frac{t_{n+1}p_n + p_{n-1}}{t_{n+1}q_n + q_{n-1}},$$

we have

$$t_{n+1}(p_n - q_n\sqrt{D}) = -p_{n-1} + q_{n-1}\sqrt{D}.$$

Thus,

$$\begin{aligned}
t_{n+1} &= t_{n+1} \cdot (-1)^{n+1}(p_n^2 - Dq_n^2) \\
&= t_{n+1}(p_n - q_n\sqrt{D}) \cdot (-1)^{n+1}(p_n + q_n\sqrt{D}) \\
&= (-p_{n-1} + q_{n-1}\sqrt{D}) \cdot (-1)^{n+1}(p_n + q_n\sqrt{D}) \\
&= (-1)^{n+1}\big((-p_np_{n-1} + q_nq_{n-1}D) + (p_nq_{n-1} - p_{n-1}q_n)\sqrt{D}\big) \\
&= (-1)^{n+1}\big((-p_np_{n-1} + q_nq_{n-1}D) + (-1)^{n-1}\sqrt{D}\big) \\
&= \sqrt{D} + (-1)^n(p_np_{n-1} - q_nq_{n-1}D).
\end{aligned}$$

In other words, we have $t_{n+1} = \sqrt{D} + C$ where C is an integer. Finally,

$$t_{n+2} = \frac{1}{t_{n+1} - \lfloor t_{n+1} \rfloor} = \frac{1}{(\sqrt{D}+C) - \lfloor \sqrt{D}+C \rfloor} = \frac{1}{\sqrt{D} - \lfloor \sqrt{D} \rfloor} = t_1,$$

thereby confirming the requested claim.

Finally, it follows from Theorem 30.9 that every solution (x,y) to $x^2 - Dy^2 = \pm 1$ with $x, y > 0$ is among $(p_{\ell m-1}, q_{\ell m-1})$ with $\ell \geq 1$. Hence, (p_{m-1}, q_{m-1}) is the fundamental solution by the monotonicity of the sequences $\{p_k\}$ and $\{q_k\}$. ■

■ **Example 30.2** **(i).** Consider $D = 13$, i.e. the equation $x^2 - 13y^2 = \pm 1$. We have $\sqrt{13} = \langle 3, \overline{1,1,1,1,6} \rangle$, which is of period $m = 5$. Also, the fourth convergent gives the fundamental solution $(x_1, y_1) = (p_4, q_4) = (18, 5)$ with $x_1^2 - 13y_1^2 = 18^2 - 13 \cdot 5^2 = (-1)^5 = -1$. Hence, we have

$$x_k = \frac{(18+5\sqrt{13})^k + (18-5\sqrt{13})^k}{2} \quad \text{and} \quad y_k = \frac{(18+5\sqrt{13})^k - (18-5\sqrt{13})^k}{2\sqrt{13}}. \tag{30.11}$$

The solutions to $x^2 - 13y^2 = 1$ are given by $(\pm x_{2k}, \pm y_{2k})$ with $k \geq 0$, e.g. $(\pm 1, 0)$, $(\pm 649, \pm 180)$, etc. The solutions to $x^2 - 13y^2 = -1$ are given by $(\pm x_{2k+1}, \pm y_{2k+1})$ with $k \geq 0$, e.g. $(\pm 18, \pm 5)$, $(\pm 23382, \pm 6485)$, etc.

(ii). Consider $D = 18$, i.e. the equation $x^2 - 18y^2 = \pm 1$. We have $\sqrt{18} = \langle 4, \overline{4,8} \rangle$, which is of period $m = 2$. Also, the first convergent gives the fundamental solution $(x_1, y_1) = (p_1, q_1) = (17, 4)$ with $x_1^2 - 18y_1^2 = 17^2 - 18 \cdot 4^2 = (-1)^2 = 1$. Hence, we have

$$x_k = \frac{(17+4\sqrt{18})^k + (17-4\sqrt{18})^k}{2} \quad \text{and} \quad y_k = \frac{(17+4\sqrt{18})^k - (17-4\sqrt{18})^k}{2\sqrt{18}}. \tag{30.12}$$

The solutions to $x^2 - 18y^2 = 1$ are given by $(\pm x_k, \pm y_k)$ with $k \geq 0$, e.g. $(\pm 1, 0)$, $(\pm 17, \pm 4)$, $(\pm 577, \pm 136)$, etc. The equation $x^2 - 18y^2 = -1$ has no solution. ■

31. Fermat's Last Theorem (I)

31.1 Fermat's Last Theorem

One of the fundamental relations in *Classical Euclidean Geometry* concerns the three sides of a right triangle, and it states that the square sum of the two shorter sides equals the square of the longest side, known as the *hypotenuse*. Arithmetically, if the two shorter sides are of length a and b, and the hypotenuse is of length c, then

$$a^2 + b^2 = c^2. \tag{31.1}$$

This relation is called the *Pythagorean Theorem*, named after the ancient Greek philosopher Pythagoras. A particularly interesting problem is to find right triangles with integer sides, and it is almost immediate to find some small examples such as $3^2 + 4^2 = 5^2$ and $5^2 + 12^2 = 13^2$. In Problem II.8 of the *Arithmetica*, the Alexandrian mathematician Diophantus formally asked for a characterization of triples of integers (a, b, c), which are now called the *Pythagorean triples*, such that (31.1) holds.

Around 1637, Fermat wrote in the margin of his copy of the *Arithmetica* the following famous and somewhat mysterious comments on the *Pythagorean triples*:

> *It is impossible to separate a cube into two cubes, or a fourth power into two fourth powers, or in general, any power higher than the second, into two like powers. I have discovered a truly marvelous proof of this, which this margin is too narrow to contain.*

Formally, Fermat claimed the following statement.

Fermat's Last Theorem For any integer $n \geq 3$, the equation

$$x^n + y^n = z^n \tag{31.2}$$

has no positive integer solutions.

R We usually call (31.2) the *Fermat equation*.

Did Fermat have a *marvelous proof*? Nobody knows but it is most likely he did *not*. Perhaps it is the power of the method of infinite descent as described in Lemma 8.4 that gave Fermat the illusion that he had one.

The initial attempts at the Fermat equation only concern specific exponents. These attempts include Fermat's own proof for the $n=4$ case based on the technique of infinite descent. The $n=3$ case is mainly attributed to Euler, while Legendre and Dirichlet independently proved Fermat's Last Theorem for $n=5$.

Here one shall note that if the Fermat equation has a positive integer solution for a certain exponent $n \geq 3$, so does the equation for any divisor d of n with $d \geq 3$. This is because if (x,y,z) is such that $x^n+y^n=z^n$, then we also have $(x^{d'})^d+(y^{d'})^d=(z^{d'})^d$ where $dd'=n$. Since Fermat has already shown the $n=4$ case, it is sufficient to prove that the Fermat equation has no positive integer solutions for $n=p$, an odd prime. Yet another observation is that if any two of x, y and z, if they exist, have a common prime factor, so does the remaining number, and hence we may cast out this common factor. Eventually, we will get x, y and z so that they are pairwise coprime.

Definition 31.1 A triple of nonzero integers (x,y,z) is called *primitive* if the numbers x, y and z are pairwise coprime.

From the above discussions, we may summarize the following fact about Fermat's Last Theorem.

Fact 31.1 To prove Fermat's Last Theorem, it is sufficient to show that the equation

$$x^p+y^p=z^p \tag{31.3}$$

has no primitive positive solutions for any odd prime number p.

The modern approaches to attack Fermat's Last Theorem exhibit more and more algebraic flavor, such as the German mathematician Ernst Kummer's development of the theory of ideals. In the 1980s, Gerhard Frey, Jean-Pierre Serre, and Ken Ribet cleverly linked the Fermat equation with elliptic curves, and showed that the validity of Fermat's Last Theorem is built on the validity of a special case of the Taniyama–Shimura–Weil conjecture which concerns the modularity for elliptic curves. Following this line, the British mathematician Andrew Wiles ultimately succeeded in proving Fermat's Last Theorem in 1994, after over three and a half centuries of waiting.

So looking back at Fermat's comments in 1637, only two thirds are completely correct — the Fermat equation has no positive integer solutions and the margin is too narrow to contain Wiles's 129-page proof (*Ann. of Math. (2)* **141** (1995), no. 3, 443–551 & 553–572; the latter is joint with his student Richard Taylor). But may Fermat really know some hidden secrets that are silently lying in *THE BOOK*?

31.2 Quadratic case: Pythagorean triples

We have seen earlier some instances of Pythagorean triples. Now our object is to characterize all of them.

Theorem 31.2 A triple of positive integers (x,y,z) is a primitive Pythagorean triple, i.e. x,y,z are pairwise coprime such that

$$x^2+y^2=z^2, \tag{31.4}$$

if and only if

$$\begin{cases} x = r^2 - s^2, \\ y = 2rs, \\ z = r^2 + s^2, \end{cases} \quad \text{or} \quad \begin{cases} x = 2rs, \\ y = r^2 - s^2, \\ z = r^2 + s^2, \end{cases} \tag{31.5}$$

where $r > s > 0$ are coprime integers of different parities.

Proof. The sufficiency is almost plain. Here we only need to show that the first case in (31.5) gives a Pythagorean triple as the second case is simply obtained by swapping x and y in the first case. First, it is straightforward that $(x,y,z) = (r^2-s^2, 2rs, r^2+s^2)$ are such that $x^2+y^2 = z^2$. Now it suffices to verify that they are pairwise coprime. Since r and s are of different parities, we find that r^2-s^2 and r^2+s^2 are odd. Hence, 2 is not a common factor of any two of r^2-s^2, $2rs$ and r^2+s^2. Assume that p is an odd prime. If p divides any two of r^2-s^2, $2rs$ and r^2+s^2, then p divides both r and s, thereby contradicting the assumption that r and s are coprime.

For necessity, we start by noting that x and y cannot be simultaneously odd, for if this is the case, then $x^2+y^2 \equiv 2 \pmod 4$, which cannot be a square. Thus, without loss of generality, we assume that x is odd and y is even so that z is odd, and shall prove that there exist coprime positive integers $r > s$ of different parities such that $x = r^2 - s^2$ and $y = 2rs$, and hence that $z = r^2+s^2$. Now we rewrite (31.4) as

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

Since we have assumed that x and z are odd, we know that $z \pm x$ are even. Let us write $z+x = 2u$ and $z-x = 2v$. Note also that $(u,v) = 1$. Otherwise, if u and v have a common prime divisor $p > 1$, then p also divides $u-v = x$ and $u+v = z$, thereby violating the assumption that (x,y,z) is primitive. Also, u and v are of different parities as x and z are supposed to be odd. Finally, we have

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

Since y is even, we find that uv is a square. Further, since $(u,v) = 1$, each of them is a square. We write $u = r^2$ and $v = s^2$. Thus, $x = u - v = r^2 - s^2$, $y = 2\sqrt{uv} = 2rs$ and $z = u+v = r^2+s^2$. Further, the assumption $r > s > 0$ comes from the fact that $x > 0$ and the assumption that $(r,s) = 1$ comes from the fact that $(u,v) = 1$. Finally, u and v have different parities and so do r and s, as required. ■

In what follows, we shall give an alternative consideration of Theorem 31.2 from a more algebraic point of view.

Let $i = \sqrt{-1}$. Recall from Sect. 26.4 that $\mathscr{O}_{\mathbb{Q}(i)} = \mathbb{Z}[i]$, which is a unique factorization domain. The units of $\mathbb{Z}[i]$ are ± 1 and $\pm i$. Also, all prime (or equivalently, irreducible) elements in $\mathbb{Z}[i]$ are:

(i) $1+i$ and its associates;
(ii) rational primes p and their associates for $p \equiv 3 \pmod 4$;
(iii) nonunit and nonassociate factors $a+bi$ of rational primes p for $p \equiv 1 \pmod 4$, i.e. $a,b \in \mathbb{Z}$ are such that $p = a^2+b^2$.

Lemma 31.3 If $x,y \in \mathbb{Z}$ of different parities are coprime in $\mathbb{Z}$, then $x+yi$ and $x-yi$ are coprime in $\mathbb{Z}[i]$.

Proof. Assume on the contrary that there is a prime element π in $\mathbb{Z}[i]$ which divides both $x+yi$ and $x-yi$. First, $\pi \nmid 2$ in $\mathbb{Z}[i]$. Otherwise, $N_{\mathbb{Q}(i)/\mathbb{Q}}(\pi) = 2$ divides $N_{\mathbb{Q}(i)/\mathbb{Q}}(x+yi) = x^2+y^2$ in $\mathbb{Z}$. But since x and y have different parities, x^2+y^2 is odd, thereby leading to a contradiction. Now $\pi \mid (x+yi)$ and $\pi \mid (x-yi)$ imply that $\pi \mid 2x$ and $\pi \mid 2yi$. Since $\pi \nmid 2$, we have $\pi \mid x$ and $\pi \mid y$. If π is associated with a rational prime $p \equiv 3 \pmod 4$, then $p \mid x$ and $p \mid y$, and this violates the assumption that x and y are coprime in $\mathbb{Z}$. If π is a factor of a rational prime $p \equiv 1 \pmod 4$, then $N_{\mathbb{Q}(i)/\mathbb{Q}}(\pi) = p$ divides $N_{\mathbb{Q}(i)/\mathbb{Q}}(x) = x^2$ and $N_{\mathbb{Q}(i)/\mathbb{Q}}(y) = y^2$ in $\mathbb{Z}$. Hence, we still have $p \mid x$ and $p \mid y$, and arrive at the same contradiction. The required claim therefore follows. ∎

Lemma 31.4 Let $\kappa, \lambda \in \mathbb{Z}[i]$ be coprime in $\mathbb{Z}[i]$. Then if $\kappa\lambda$ is associated with a square in $\mathbb{Z}[i]$, so are κ and λ.

Proof. We uniquely factor $\kappa\lambda$ as $\kappa\lambda = u\pi_1^{2\alpha_1}\cdots\pi_k^{2\alpha_k}$, where u is a unit, and $\pi_1, \ldots, \pi_k$ are distinct prime elements. In particular, the powers $2\alpha_1, \ldots, 2\alpha_k$ are even since $\kappa\lambda$ is associated with a square. Now since κ and λ are coprime, each $\pi_j^{2\alpha_j}$ is exclusively in one of the factorizations of κ and λ, thereby implying the required result. ∎

Now we present the second proof of Theorem 31.2.

Second Proof of Theorem 31.2. Here we only establish the necessity. Recall that (x,y,z) is a primitive Pythagorean triple so that $x^2+y^2=z^2$. As we have argued earlier, x and y have different parities and without loss of generality, we assume that x is odd and y is even. Now by Lemma 31.3, $x+yi$ and $x-yi$ are coprime in $\mathbb{Z}[i]$. Note that

$$z^2 = (x+yi)(x-yi).$$

By Lemma 31.4, we may write

$$x+yi = u(r+si)^2,$$

where $u \in \{\pm 1, \pm i\}$ is a unit and $r,s \in \mathbb{Z}$. Since $x,y>0$, neither of r and s are 0. We may further assume that $r,s>0$ as in other cases we may factor out a unit from $r+si$. Therefore,

$$x+yi = u\big((r^2-s^2)+2rsi\big).$$

Since x is assumed to be odd and y is assumed to be even, then r and s have different parities and $u \in \{\pm 1\}$. Further, $y>0$ and $r,s>0$ imply that $u=1$. Thus, $(x,y) = (r^2-s^2, 2rs)$ so that $z = r^2+s^2$ while we additionally require that $r>s$ so that $x>0$. Finally, r and s must be coprime to ensure that (x,y,z) is primitive. ∎

31.3 Quartic case: An elementary approach

Now we shall apply Theorem 31.2 to prove the quartic case of Fermat's Last Theorem and the proof is essentially built on Fermat's method of infinite descent. In fact, we establish the following stronger result.

Theorem 31.5 The equation

$$x^4+y^4=z^2 \tag{31.6}$$

has no positive integer solutions.

Proof. Suppose on the contrary that (31.6) has a positive integer solution and we may further assume that $(x,y)=1$ by casting out all common factors of x and y. Thus, the triple (x,y,z) is primitive. Further, x and y cannot be simultaneously odd as if this is the case, then $z^2 \equiv 2 \pmod 4$, which is impossible. Hence, we assume that x is odd and y is even, so that z is odd.

Let (X,Y,Z) be such a primitive solution with Z minimal. If we rewrite (31.6) as

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

then by Theorem 31.2, there exist coprime integers $r>s>0$ of different parities such that

$$X^2=r^2-s^2, \qquad Y^2=2rs, \qquad Z=r^2+s^2.$$

Note that if r is even and s is odd, then $X^2 \equiv -1 \pmod 4$, which is impossible. Hence, r is odd and s is even, and we write $s=2t$. Thus, $Y^2=4rt$. Since Y is even and $(r,t)=1$, we may write

$$r=k^2, \qquad t=\ell^2,$$

where $k,\ell>0$ are integers with $(k,\ell)=1$. In particular, since r is odd, so is k. It follows from $X^2=r^2-s^2=r^2-(2t)^2$ that

$$X^2+(2\ell^2)^2=(k^2)^2.$$

Applying Theorem 31.2 again to the above, we have coprime integers $a>b>0$ of different parities such that

$$X=a^2-b^2, \qquad 2\ell^2=2ab, \qquad k^2=a^2+b^2.$$

Since $ab=\ell^2$ and $(a,b)=1$, we may further write

$$a=c^2, \qquad b=d^2,$$

so that

$$k^2=c^4+d^4.$$

Note that here c and d are coprime and of different parities. By renaming c and d, we may further assume that c is odd and d is even. The above relation gives another solution to (31.6), namely, $(x,y,z)=(c,d,k)$. However, we have

$$k \le k^2 = r \le r^2 < r^2+s^2 = Z,$$

thereby contradicting the minimality of Z. ■

31.4 Quartic case: An algebraic approach

Let us further consider (31.6) in $\mathbb{Z}[i]$. It turns out that even if the domain of solutions is substantially extended, we still have merely trivial solutions in the sense that at least one of x, y and z is zero. In this section, we will establish this result following the idea of David Hilbert (*Jahresber. Dtsch. Math.-Ver.* **4** (1897), 175–546).

Throughout, all arithmetic is done in $\mathbb{Z}[i]$.

Definition 31.2 Let $\xi,\eta,\theta \in \mathbb{Z}[i]$ be nonzero. We say the triple (ξ,η,θ) is *primitive* if ξ, η and θ are pairwise coprime.

Note that for (ξ,η,θ) with

$$\xi^4+\eta^4=\theta^2, \tag{31.7}$$

if any prime element divides two of ξ, η and θ, it divides the remaining number. Hence, we may cast out all common factors and it is sufficient to consider primitive solutions.

Let $\lambda = 1-i$ in this section. Then λ is an associate of $1+i$, and thus a prime element in $\mathbb{Z}[i]$. Further, $\lambda^2=-2i$ is an associate of 2, $\lambda^4=-4$ is an associate of 4 and $\lambda^6=8i$ is an associate of 8.

Lemma 31.6 Let $\alpha \in \mathbb{Z}[i]$ be such that $\lambda \nmid \alpha$. Then

$$\alpha^2 \equiv \pm 1 \pmod{\lambda^4}, \tag{31.8}$$
$$\alpha^4 \equiv 1 \pmod{\lambda^6}. \tag{31.9}$$

Proof. Note that for any number κ in $\mathbb{Z}[i]$, there are four possibilities modulo 2, namely $\kappa \equiv 0,1,i,\lambda \pmod 2$. Since $\lambda \mid 2$, we know that for α with $\lambda \nmid \alpha$, it is only possible that $\alpha \equiv 1,i \pmod 2$. If $\alpha = 2\delta+1$ with $\delta \in \mathbb{Z}[i]$, then $\alpha^2=4\delta^2+4\delta+1$, implying that $\alpha^2 \equiv 1 \pmod 4$, and equivalently that $\alpha^2 \equiv 1 \pmod{\lambda^4}$. Similarly, if $\alpha = 2\delta+i$, then $\alpha^2 \equiv -1 \pmod{\lambda^4}$. By the same reasoning, we further have $\alpha^4 \equiv 1 \pmod{\lambda^6}$ whenever $\lambda \nmid \alpha$. ■

We consider a variant of (31.7).

Lemma 31.7 If there exist a unit u of $\mathbb{Z}[i]$ and a primitive triple (ξ,η,θ) in $\mathbb{Z}[i]$ with $\lambda \mid \xi$ such that

$$u\xi^4+\eta^4=\theta^2,$$

then we must have $\lambda^2 \mid \xi$.

Proof. Since (ξ,η,θ) is primitive while $\lambda \mid \xi$, we have $\lambda \nmid \eta$ and $\lambda \nmid \theta$. Then $\eta^4 \equiv 1 \pmod{\lambda^6}$ and thus $\theta^2 = u\xi^4+\eta^4 \equiv 0+1=1 \pmod{\lambda^4}$. We conclude that $\theta \equiv 1 \pmod{\lambda^2}$ as $\lambda \nmid \theta$ implies that $\theta \equiv 1,i \pmod{\lambda^2}$ but for the latter case we further have $\theta^2 \equiv -1 \pmod{\lambda^4}$, which is not true. Let us write $\theta = \lambda^2\delta+1$ with $\delta \in \mathbb{Z}[i]$. If $\delta \equiv 0,\lambda \pmod 2$, then $\lambda \mid \delta$; if $\delta \equiv 1,i \pmod 2$, then $\delta+i \equiv \lambda,0 \pmod 2$ so that $\lambda \mid (\delta+i)$. Hence, $\lambda \mid \delta(\delta+i)$. Since

$$\theta^2-1=(\theta-1)(\theta+1)=\lambda^2\delta(\lambda^2\delta+2)=\lambda^2\delta(\lambda^2\delta+\lambda^2 i)=\lambda^4\delta(\delta+i),$$

we have $\theta^2 \equiv 1 \pmod{\lambda^5}$. Therefore, $u\xi^4=\theta^2-\eta^4 \equiv 1-1=0 \pmod{\lambda^5}$, which further implies that $\lambda^2 \mid \xi$. ■

Now we require a subtle application of the method of infinite descent.

Theorem 31.8 For any unit u of $\mathbb{Z}[i]$, there is no primitive triple (ξ,η,θ) in $\mathbb{Z}[i]$ with $\lambda \mid \xi$ such that

$$u\xi^4+\eta^4=\theta^2. \tag{31.10}$$

Proof. The core of this proof relies on the claim that whenever we are given a unit u and a primitive triple (ξ,η,θ) with $\lambda^n \parallel \xi$, i.e. $\lambda^n \mid \xi$ and $\lambda^{n+1} \nmid \xi$, for some $n \geq 2$, such that

$u\xi^4+\eta^4=\theta^2$ holds, we are always able to find another unit u' and another primitive triple (ξ',η',θ') with $\lambda^{n-1}\parallel\xi'$ such that $u'\xi'^4+\eta'^4=\theta'^2$.

Now suppose on the contrary to the lemma that there exist a unit u_1 and a primitive triple (ξ_1,η_1,θ_1) with $\lambda\mid\xi_1$ such that $u_1\xi_1^4+\eta_1^4=\theta_1^2$. Then by the above claim together with the technique of infinite descent, we are ultimately led to a unit u_0 and a primitive triple (ξ_0,η_0,θ_0) with $\lambda\parallel\xi_0$ such that $u_0\xi_0^4+\eta_0^4=\theta_0^2$. But this contradicts Lemma 31.7 as $\lambda^2\nmid\xi_0$. Hence, the desired result is true.

From now on, we prove the initial claim. Again, we have $\lambda\nmid\eta$ and $\lambda\nmid\theta$. Note that

$$u\xi^4=(\theta+\eta^2)(\theta-\eta^2).$$

Since $\lambda^2\mid\lambda^n\parallel\xi$, we see that $\lambda^8\mid u\xi^4=(\theta+\eta^2)(\theta-\eta^2)$. It follows that λ^2 divides at least one of $\theta+\eta^2$ and $\theta-\eta^2$ as λ is a prime element. If $\theta\pm\eta^2\equiv 0 \pmod{\lambda^2}$, then $\theta\mp\eta^2=(\theta\pm\eta^2)\mp 2\eta^2\equiv 0\mp 0=0 \pmod{\lambda^2}$ since we also have $\lambda^2\mid 2$. Thus, λ^2 divides both $\theta+\eta^2$ and $\theta-\eta^2$. If there exists an additional prime element π such that $\lambda^2\pi$ divides both $\theta+\eta^2$ and $\theta-\eta^2$, then $\lambda^2\pi\mid 2\theta$ and $\lambda^2\pi\mid 2\eta^2$, so that $\pi\mid\theta$ and $\pi\mid\eta^2$ while the latter further gives $\pi\mid\eta$. But this violates the assumption that η and θ are coprime. Hence, $(\theta+\eta^2,\theta-\eta^2)=\lambda^2$.

Let us write

$$\begin{cases}\theta\pm\eta^2=\lambda^2\delta_1,\\ \theta\mp\eta^2=\lambda^2\delta_2,\end{cases}$$

where $\delta_1,\delta_2\in\mathbb{Z}[i]$ are coprime. Then from

$$u\xi^4=\lambda^4\delta_1\delta_2,$$

we may further write $\delta_1=v_1\kappa_1^4$ and $\delta_2=v_2\kappa_2^4$ where $\kappa_1,\kappa_2\in\mathbb{Z}[i]$ are coprime and v_1,v_2 are units. Meanwhile, since $\lambda^{4n}\parallel\xi^4$, we have $\lambda^{4(n-1)}\parallel\delta_1\delta_2=v_1v_2\kappa_1^4\kappa_2^4$. As κ_1 and κ_2 are coprime, we assume without loss of generality that $\lambda^{n-1}\parallel\kappa_1$ and $\lambda\nmid\kappa_2$.

Noting that

$$\pm 2\eta^2=\lambda^2\delta_1-\lambda^2\delta_2=-2i(v_1\kappa_1^4-v_2\kappa_2^4),$$

we have

$$\eta^2=w_1\kappa_1^4+w_2\kappa_2^4,$$

where we put $w_1=\mp iv_1$ and $w_2=\pm iv_2$, both of which are units. Recalling that $n\geq 2$ and hence that $\lambda\mid\lambda^{n-1}\parallel\kappa_1$, we get

$$\eta^2\equiv w_2\kappa_2^4 \pmod{\lambda^4}.$$

However, since $\lambda\nmid\eta$ and $\lambda\nmid\kappa_2$, Lemma 31.6 tells us that $\eta^2\equiv\pm 1 \pmod{\lambda^4}$ and $\kappa_2^4\equiv 1 \pmod{\lambda^4}$. Hence, $w_2\equiv\pm 1 \pmod{\lambda^4}$. However, w_2 is a unit, so $w_2=\pm 1$. If $w_2=1$, we choose

$$u'=w_1,\qquad(\xi',\eta',\theta')=(\kappa_1,\kappa_2,\eta);$$

if $w_2=-1$, we choose

$$u'=-w_1,\qquad(\xi',\eta',\theta')=(\kappa_1,\kappa_2,i\eta).$$

In both cases, we have $u'\xi'^4+\eta'^4=\theta'^2$ where u' is a unit, (ξ',η',θ') is primitive, and $\lambda^{n-1}\parallel\xi'$, as requested. ■

Finally, the nonexistence of primitive solutions to $\xi^4+\eta^4=\theta^2$ is an immediate implication of the above result.

Theorem 31.9 The equation

$$\xi^4 + \eta^4 = \theta^2$$

has no primitive solutions in $\mathbb{Z}[i]$.

Proof. Assume on the contrary that there is a primitive solution (ξ, η, θ).

We first prove that λ divides exactly one of ξ and η. If this is not the case, i.e. λ divides neither of them, then $\xi^4 \equiv 1 \pmod{\lambda^6}$ and $\eta^4 \equiv 1 \pmod{\lambda^6}$ by Lemma 31.6, thereby implying that $\theta^2 \equiv 1 + 1 = 2 = i\lambda^2 \pmod{\lambda^6}$. Therefore, $\lambda^2 \mid \theta^2$ so that $\lambda \mid \theta$ since λ is a prime element. We further note from $4 \mid \lambda^6$ that $\theta^2 \equiv 2 \not\equiv 0 \pmod 4$, i.e. $4 \nmid \theta^2$, or equivalently, $\lambda^4 \nmid \theta^2$. Hence, $\lambda^2 \nmid \theta$. Now let us write $\theta = \lambda\theta'$, where $\lambda \nmid \theta'$. Since $\lambda^2\theta'^2 = \theta^2 \equiv i\lambda^2 \pmod{\lambda^6}$, we have $\theta'^2 \equiv i \pmod{\lambda^4}$. However, by Lemma 31.6, it is only possible that $\theta'^2 \equiv \pm 1 \pmod{\lambda^4}$ as $\lambda \nmid \theta'$. We arrive at a contradiction.

Without loss of generality suppose that $\lambda \mid \xi$. Then we are led to an instance of (31.10) with $u = 1$, a unit of $\mathbb{Z}[i]$. But such an instance should not exist by Theorem 31.8. ■

32. Fermat's Last Theorem (II)

32.1 Cubic case: An algebraic approach

As we have seen in the previous lecture, for the quartic case, the algebraic approach seems to be much more complicated than the elementary one. However, there is still an advantage that it may be naturally transplanted to the cubic case but with a focus on $\mathscr{O}_{\mathbb{Q}(\sqrt{-3})}$. However, to provide an elementary proof, one should be extremely careful for even as legendary as Euler would miss some crucial steps.

In Sect. 26.5, we considered $\mathscr{O}_{\mathbb{Q}(\sqrt{-3})}$ as $\mathbb{Z}[\frac{1+\sqrt{-3}}{2}]$. Here, we shall use a slightly different generator, namely, $\mathscr{O}_{\mathbb{Q}(\sqrt{-3})} = \mathbb{Z}[\frac{-1+\sqrt{-3}}{2}]$, for the sake of computational convenience.

Throughout, let $\zeta = \frac{-1+\sqrt{-3}}{2}$ for brevity. Then $\zeta^2 = \frac{-1-\sqrt{-3}}{2}$, $\zeta^3 = 1$ and $1+\zeta+\zeta^2 = 0$. It is already known that $\mathscr{O}_{\mathbb{Q}(\sqrt{-3})} = \mathbb{Z}[\zeta]$ is a unique factorization domain. The units of $\mathbb{Z}[\zeta]$ are ± 1 and $\frac{\pm 1 \pm \sqrt{-3}}{2}$. Also, all prime (or equivalently, irreducible) elements in $\mathbb{Z}[\zeta]$ are:

(i) $\sqrt{-3}$ and its associates;

(ii) rational primes p and their associates for $p = 2$ or $p \equiv 5 \pmod 6$;

(iii) nonunit and nonassociate factors $a + b\zeta$ of rational primes p for $p \equiv 1 \pmod 6$, i.e. $a, b \in \mathbb{Z}$ are such that $p = a^2 - ab + b^2$.

Throughout, all arithmetic is done in $\mathbb{Z}[\zeta]$.

Let $\lambda = \frac{3-\sqrt{-3}}{2}$ in this section. Then $\lambda = 1 - \zeta = \sqrt{-3} \cdot \frac{-1-\sqrt{-3}}{2}$ is an associate of $\sqrt{-3}$, and thus a prime element in $\mathbb{Z}[\zeta]$. Also, $\lambda^2 = -3\zeta^{-2}$ is an associate of 3.

Lemma 32.1 Let $\alpha \in \mathbb{Z}[\zeta]$ be such that $\lambda \nmid \alpha$. Then

$$\alpha^3 \equiv \pm 1 \pmod{\lambda^4}. \tag{32.1}$$

Proof. Note that for any number κ in $\mathbb{Z}[\zeta]$, there are three possibilities modulo λ, namely $\kappa \equiv 0, \pm 1 \pmod{\lambda}$ for $\mathbb{Z}[\zeta] = \mathbb{Z}[1-\zeta] = \mathbb{Z}[\lambda]$ and $3 = -\zeta^2\lambda^2 \equiv 0 \pmod{\lambda}$. Hence for α with $\lambda \nmid \alpha$, it is only possible that $\alpha \equiv \pm 1 \pmod{\lambda}$. If $\alpha = \lambda\delta + 1$ with $\delta \in \mathbb{Z}[i]$, then

$$\alpha^3 = (\lambda\delta + 1)^3 = \lambda^3\delta^3 + 3\lambda^2\delta^2 + 3\lambda\delta + 1$$

$$
\begin{aligned}
&= \lambda^3\delta^3 - \zeta^2\lambda^4\delta^2 - \zeta^2\lambda^3\delta + 1\\
&\equiv \lambda^3\delta^3 - \zeta^2\lambda^3\delta + 1\\
&= \lambda^3\delta(\delta+\zeta)(\delta-\zeta)+1\\
&= \lambda^3\delta\big(\delta+(1-\lambda)\big)\big(\delta-(1-\lambda)\big)+1\\
&\equiv \lambda^3\delta(\delta+1)(\delta-1)+1 \pmod{\lambda^4}.
\end{aligned}
$$

Since $\delta \equiv 0, \pm 1 \pmod{\lambda}$, we have $\delta(\delta+1)(\delta-1) \equiv 0 \pmod{\lambda}$. Thus, $\alpha^3 = (\lambda\delta+1)^3 \equiv 1 \pmod{\lambda^4}$. If $\alpha = \lambda\delta - 1$, then $-\alpha = -\lambda\delta + 1$, so that $\alpha^3 = -(-\alpha)^3 = -(-\lambda\delta+1)^3 \equiv -1 \pmod{\lambda^4}$. ∎

Definition 32.1 Let $\xi, \eta, \theta \in \mathbb{Z}[\zeta]$ be nonzero. We say the triple (ξ, η, θ) is *primitive* if ξ, η and θ are pairwise coprime.

Note that for (ξ, η, θ) with

$$\xi^3 + \eta^3 = \theta^3, \tag{32.2}$$

if any prime element divides two of ξ, η and θ, it divides the remaining number. Hence, we may cast out all common factors and merely consider primitive solutions. Further, since $-\theta^3 = (-\theta)^3$, it is equivalent to consider

$$\xi^3 + \eta^3 + \theta^3 = 0. \tag{32.3}$$

We also start with a variant of (32.3).

Lemma 32.2 If there exist a unit u of $\mathbb{Z}[\zeta]$ and a primitive triple (ξ, η, θ) in $\mathbb{Z}[\zeta]$ with $\lambda \mid \xi$ such that

$$u\xi^3 + \eta^3 + \theta^3 = 0,$$

then we must have $\lambda^2 \mid \xi$.

Proof. Since (ξ, η, θ) is primitive while $\lambda \mid \xi$, we have $\lambda \nmid \eta$ and $\lambda \nmid \theta$. Hence, $u\xi^3 = -\eta^3 - \theta^3 \equiv \pm 1 \pm 1 = 0$ or $\pm 2 \pmod{\lambda^4}$. Further, since $\lambda \mid u\xi^3$ while $\lambda \nmid (\pm 2)$, we must have $u\xi^3 \equiv 0 \pmod{\lambda^4}$. This implies that $\lambda^2 \mid \xi$. ∎

The following is again a consequence of the technique of infinite descent.

Theorem 32.3 For any unit u of $\mathbb{Z}[\zeta]$, there is no primitive triple (ξ, η, θ) in $\mathbb{Z}[\zeta]$ with $\lambda \mid \xi$ such that

$$u\xi^3 + \eta^3 + \theta^3 = 0. \tag{32.4}$$

Proof. We shall show that whenever we are given a unit u and a primitive triple (ξ, η, θ) with $\lambda^n \parallel \xi$ for some $n \geq 2$, such that $u\xi^3 + \eta^3 + \theta^3 = 0$ holds, we are always able to find another unit u' and another primitive triple (ξ', η', θ') with $\lambda^{n-1} \parallel \xi'$ such that $u'\xi'^3 + \eta'^3 + \theta'^3 = 0$.

Now suppose on the contrary to the lemma that there exist a unit u_1 and a primitive triple $(\xi_1, \eta_1, \theta_1)$ with $\lambda \mid \xi_1$ such that $u_1\xi_1^3 + \eta_1^3 + \theta_1^3 = 0$. Then the method of infinite descent tells us that there must exist a unit u_0 and a primitive triple $(\xi_0, \eta_0, \theta_0)$ with $\lambda \parallel \xi_0$ such that $u_0\xi_0^3 + \eta_0^3 + \theta_0^3 = 0$, which is however impossible by Lemma 32.2. Hence, the desired result is true.

From now on, we prove the initial claim. Again, we have $\lambda \nmid \eta$ and $\lambda \nmid \theta$. Note that

$$-u\xi^3 = (\theta+\eta)(\theta+\zeta\eta)(\theta+\zeta^2\eta).$$

Since $\lambda \mid \xi$, we see that λ divides at least one of $\theta+\eta$, $\theta+\zeta\eta$ and $\theta+\zeta^2\eta$ as λ is a prime element. Further, it follows from $\zeta = 1-\lambda \equiv 1 \pmod{\lambda}$ that $\theta+\eta \equiv \theta+\zeta\eta \equiv \theta+\zeta^2\eta \pmod{\lambda}$. Thus, λ divides all of $\theta+\eta$, $\theta+\zeta\eta$ and $\theta+\zeta^2\eta$. If there exists an additional prime element π such that $\lambda\pi$ divides both $\theta+\eta$ and $\theta+\zeta\eta$, then $\lambda\pi \mid \lambda\eta$ and $\lambda\pi \mid \lambda\theta$ so that $\pi \mid \eta$ and $\pi \mid \theta$, but this violates the assumption that η and θ are coprime. By the same reasoning with an extra application of the fact that $\zeta^3 = 1$ when treating $\theta+\eta$ and $\theta+\zeta^2\eta$, we find that

$$(\theta+\eta,\theta+\zeta\eta) = (\theta+\zeta\eta,\theta+\zeta^2\eta) = (\theta+\zeta^2\eta,\theta+\eta) = \lambda.$$

Let us write

$$\begin{cases} \theta+\eta = \lambda\delta_1, \\ \theta+\zeta\eta = \lambda\delta_2, \\ \theta+\zeta^2\eta = \lambda\delta_3, \end{cases}$$

where $\delta_1,\delta_2,\delta_3 \in \mathbb{Z}[\zeta]$ are pairwise coprime. Then from

$$-u\xi^3 = \lambda^3\delta_1\delta_2\delta_3,$$

we may further write $\delta_1 = v_1\kappa_1^3$, $\delta_2 = v_2\kappa_2^3$ and $\delta_3 = v_3\kappa_3^3$ where $\kappa_1,\kappa_2,\kappa_3 \in \mathbb{Z}[\zeta]$ are pairwise coprime and v_1,v_2,v_3 are units. Meanwhile, since $\lambda^{3n} \parallel \xi^3$, we have $\lambda^{3(n-1)} \parallel \delta_1\delta_2\delta_3 = v_1v_2v_3\kappa_1^3\kappa_2^3\kappa_3^3$. As κ_1, κ_2 and κ_3 are pairwise coprime, we assume without loss of generality that $\lambda^{n-1} \parallel \kappa_1$, $\lambda \nmid \kappa_2$ and $\lambda \nmid \kappa_3$; other cases may be understood by replacing η with $\zeta\eta$ and $\zeta^2\eta$.

Noting that

$$0 = (\theta+\eta)+\zeta(\theta+\zeta\eta)+\zeta^2(\theta+\zeta^2\eta),$$

we have

$$\begin{aligned} 0 &= v_1\kappa_1^3+\zeta v_2\kappa_2^3+\zeta^2 v_3\kappa_3^3 \\ &= w_1\kappa_1^3+w_2\kappa_2^3+\kappa_3^3, \end{aligned}$$

where we put $w_1 = \zeta^{-2}v_1v_3^{-1}$ and $w_2 = \zeta^{-1}v_2v_3^{-1}$, both of which are units. Recalling that $n \geq 2$ and hence that $\lambda \mid \lambda^{n-1} \parallel \kappa_1$, we get

$$0 \equiv w_2\kappa_2^3+\kappa_3^3 \pmod{\lambda^3}.$$

However, since $\lambda \nmid \kappa_2$ and $\lambda \nmid \kappa_3$, Lemma 32.1 tells us that $\kappa_2^3 \equiv \pm 1 \pmod{\lambda^4}$ and $\kappa_3^3 \equiv \pm 1 \pmod{\lambda^4}$. Hence, $w_2 \equiv \pm 1 \pmod{\lambda^3}$. However, w_2 is a unit, so $w_2 = \pm 1$. If $w_2 = 1$, we choose

$$u' = w_1, \qquad (\xi',\eta',\theta') = (\kappa_1,\kappa_2,\kappa_3);$$

if $w_2 = -1$, we choose

$$u' = w_1, \qquad (\xi',\eta',\theta') = (\kappa_1,\tfrac{1+\sqrt{-3}}{2}\kappa_2,\kappa_3).$$

In both cases, we have $u'\xi'^3+\eta'^3+\theta'^3 = 0$ where u' is a unit, (ξ',η',θ') is primitive, and $\lambda^{n-1} \parallel \xi'$, as requested. ■

Finally, we establish the nonexistence of primitive solutions to $\xi^3+\eta^3+\theta^3 = 0$.

Theorem 32.4 The equation

$$\xi^3+\eta^3+\theta^3=0$$

has no primitive solutions in $\mathbb{Z}[\zeta]$.

Proof. Assume on the contrary that there is a primitive solution (ξ,η,θ).

We first prove that λ divides exactly one of ξ, η and θ. If this is not the case, i.e. λ divides none of them, then $\xi^3\equiv\pm1 \pmod{\lambda^4}$, $\eta^3\equiv\pm1 \pmod{\lambda^4}$ and $\theta^3\equiv\pm1 \pmod{\lambda^4}$ by Lemma 32.1, thereby implying that $0\equiv\pm1\pm1\pm1\equiv\pm1$ or $\pm3 \pmod{\lambda^4}$. But both cases are impossible, and we arrive at a contradiction.

Without loss of generality suppose that $\lambda\mid\xi$. Then we are led to an instance of (32.4) with $u=1$, a unit of $\mathbb{Z}[\zeta]$. But such an instance should not exist by Theorem 32.3. ■

32.2 Cubic case: An elementary approach

We start with an analog of Lemma 8.3.

Lemma 32.5 Let $x_1,y_1,x_2,y_2\in\mathbb{R}$. Then

$$(x_1^2+3y_1^2)(x_2^2+3y_2^2)=(x_1x_2-3y_1y_2)^2+3(x_1y_2+y_1x_2)^2. \tag{32.5}$$

Proof. We may either verify by a direct calculation or make use of the fact that

$$N_{\mathbb{Q}(\sqrt{-3})/\mathbb{Q}}(\alpha)N_{\mathbb{Q}(\sqrt{-3})/\mathbb{Q}}(\beta)=N_{\mathbb{Q}(\sqrt{-3})/\mathbb{Q}}(\alpha\beta)$$

where $\alpha=x_1+y_1\sqrt{-3}$ and $\beta=x_2+y_2\sqrt{-3}$. ■

Using Lemma 32.5 twice, we see that if m can be represented as

$$m=r^2+3s^2, \tag{32.6}$$

then

$$m^3=a^2+3b^2, \tag{32.7}$$

where

$$\begin{cases} a=r(r+3s)(r-3s),\\ b=3s(r+s)(r-s). \end{cases} \tag{32.8}$$

Now, a crucial question is that whenever we write m^3 as in (32.7), possibly with suitable restrictions to a and b, are there always integers r and s such that (32.6) and (32.8) hold? Such an argument was missing in Euler's original elementary proof of the cubic case of Fermat's Last Theorem presented in his 1770 book *Vollständige Anleitung zur Algebra*. However, this gap was not explicitly pointed out until 1894 by J. Schumacher from Göppingen (*Z. Math. Naturwiss. Unterricht* **25** (1894), 350–351).

In what follows we shall adopt a neat reasoning due to Stan Dolan (*Math. Gaz.* **96** (2012), no. 535, 99–102).

Lemma 32.6 Let X and Y be coprime integers with $N=X^2+3Y^2$. Then either $N\equiv1 \pmod 2$ or $N\equiv4 \pmod 8$.

Proof. Since X and Y are coprime, either they are simultaneously odd or they have different parities. For the former case, we have $X^2 \equiv Y^2 \equiv 1 \pmod 8$ so that $N \equiv 4 \pmod 8$. For the latter case, we immediately see that N is odd. ■

Lemma 32.7 Let X and Y be coprime integers with $N = X^2 + 3Y^2$. Let x_1 and y_1 be coprime integers with $n_1 = x_1^2 + 3y_1^2$ such that $n_1 \mid N$ and $n_1 \mid (x_1Y - y_1X)$. Then there exist coprime integers x_2 and y_2 with $n_2 = x_2^2 + 3y_2^2$ such that the following relations hold:

(i) $N = n_1n_2$;
(ii) $X = x_1x_2 - 3y_1y_2$;
(iii) $Y = x_1y_2 + y_1x_2$;
(iv) $n_2 \mid (x_2Y - y_2X)$.

Proof. We shall show that $(x_2, y_2) = (\frac{x_1X+3y_1Y}{n_1}, \frac{x_1Y-y_1X}{n_1})$ is the desired choice. First, Parts (i), (ii) and (iii) can be verified by direct calculations. Further, n_2 and y_2 are integers since $n_1 \mid N$ and $n_1 \mid (x_1Y - y_1X)$. Hence x_2 is also an integer as $x_2^2 = n_2 - 3y_2^2$. Now, if $d = (x_2, y_2)$, then d divides both X and Y. But since we have assumed that X and Y are coprime, we know that $d = 1$ and hence that x_2 and y_2 are coprime. Finally, we compute that $x_2Y - y_2X = y_1n_2$ so that Part (iv) is true. ■

The following argument is the most important.

Theorem 32.8 Let X and Y be coprime integers with $N = X^2 + 3Y^2$. Then for any factorization $N = n_1n_2$ where n_1 and n_2 are not simultaneously even, we can find coprime integers x_1 and y_1 with $n_1 = x_1^2 + 3y_1^2$ and coprime integers x_2 and y_2 with $n_2 = x_2^2 + 3y_2^2$ such that

(i) $X = x_1x_2 - 3y_1y_2$;
(ii) $Y = x_1y_2 + y_1x_2$;
(iii) $n_1 \mid (x_1Y - y_1X)$;
(iv) $n_2 \mid (x_2Y - y_2X)$.

Proof. We argue by induction on $N = X^2 + 3Y^2$. When $N = 1$, i.e. $N = (\pm 1)^2 + 3 \cdot 0^2$, then for the factorization $1 = 1 \cdot 1$, we may choose the pairs $(x_1, y_1) = (1, 0)$ and $(x_2, y_2) = (\pm 1, 0)$ so that the required conditions are satisfied.

Supposing that the statement is true for $1, \ldots, N-1$ where $N \geq 2$, we prove the statement for N. If N cannot be written as $N = X^2 + 3Y^2$ with X and Y coprime, then we are done. Otherwise, let X and Y be arbitrary coprime integers such that $N = X^2 + 3Y^2$. Consider an arbitrary factorization $N = n_1n_2$, and without loss of generality, let $n_1 \leq n_2$.

Recalling that n_1 and n_2 are not simultaneously even, it follows from Lemma 32.6 that n_1 is of the form $2k+1$ or $4(2k+1)$. Also, when $n_1 = 1$ so that $n_2 = N$, we may find pairs $(x_1, y_1) = (1, 0)$ and $(x_2, y_2) = (X, Y)$ satisfying the required conditions. Below, we assume that $n_1 \geq 3$.

Note that n_1 and Y are coprime. This is because if there is a prime dividing both n_1 and Y, it also divides $n_1n_2 - 3Y^2 = N - 3Y^2 = X^2$, and thus X. But X and Y are assumed to be coprime. Let X_1 be such that $X_1Y \equiv X \pmod{n_1}$; this X_1 must exist as $(n_1, Y) = 1$ and $n_1 \geq 3$. We further choose X_1 so that $-\frac{n_1}{2} < X_1 \leq \frac{n_1}{2}$. Since $X^2 + 3Y^2 = N = n_1n_2$, we have $X^2 + 3Y^2 \equiv 0 \pmod{n_1}$ so that $(XY^{-1})^2 \equiv -3 \pmod{n_1}$. Now, $X_1^2 + 3 \equiv (XY^{-1})^2 + 3 \equiv 0 \pmod{n_1}$.

Define $N_1 = X_1^2 + 3 = X_1^2 + 3 \cdot 1^2$. From the above argument, we know that $n_1 \mid N_1$ so we may write $N_1 = n_1n_3$. By Lemma 32.6, N_1 is of the form $2k+1$ or $4(2k+1)$, while so is n_1 as argued earlier. Hence, n_1 and n_3 are not simultaneously even. Now we note that

$N_1 = X_1^2 + 3 \le (\frac{n_1}{2})^2 + 3 < n_1^2 \le n_1 n_2 = N$. By the inductive hypothesis, there must exist pairs (x_1, y_1) and (x_3, y_3) such that the required conditions hold. In particular, we have **(a).** x_1 and y_1 are coprime with $n_1 = x_1^2 + 3y_1^2$; **(b).** $n_1 \mid (x_1 \cdot 1 - y_1 X_1)$, which implies that $0 \equiv x_1 - y_1 X_1 \equiv x_1 Y - y_1 X_1 Y \equiv x_1 Y - y_1 X \pmod{n_1}$, i.e. $n_1 \mid (x_1 Y - y_1 X)$.

Recall further that $N = n_1 n_2$ so that $n_1 \mid N$. Then by Lemma 32.7, there exist coprime integers x_2 and y_2 with $n_2 = x_2^2 + 3y_2^2$ such that $X = x_1 x_2 - 3y_1 y_2$, $Y = x_1 y_2 + y_1 x_2$ and $n_2 \mid (x_2 Y - y_2 X)$. Thus, the pairs (x_1, y_1) and (x_2, y_2) are as desired. ■

The above theorem immediately gives the missing justification of Euler.

Corollary 32.9 Let a and b be coprime integers such that $a^2 + 3b^2 = m^3$. Then there exist coprime integers r and s with $r^2 + 3s^2 = m$ such that

$$\begin{cases} a = r(r+3s)(r-3s), \\ b = 3s(r+s)(r-s). \end{cases}$$

Proof. Noting that a cube cannot be congruent to 4 modulo 8, by Lemma 32.6, m^3 is odd, and so is m. Considering the factorization $m^3 = m \cdot m^2$ and applying Theorem 32.8, we find coprime integers r_0 and s_0 with $m = r_0^2 + 3s_0^2$ and coprime integers a_2 and b_2 with $m^2 = a_2^2 + 3b_2^2$ such that

$$a = r_0 a_2 - 3s_0 b_2, \qquad b = r_0 b_2 + s_0 a_2,$$

together with $m \mid (r_0 b - s_0 a)$ and $m^2 \mid (a_2 b - b_2 a)$. Further, the two divisibility properties imply that $0 \equiv a_2(r_0 b - s_0 a) - r_0(a_2 b - b_2 a) = a(r_0 b_2 - s_0 a_2) \pmod{m}$, i.e. $m \mid a(r_0 b_2 - s_0 a_2)$. Note that if there is a prime p dividing both a and m, then p^2 divides $m^3 - a^2 = 3b^2$, which gives that $p \mid b$. However, this violates the assumption that a and b are coprime. Hence, $(m, a) = 1$, so that $m \mid (r_0 b_2 - s_0 a_2)$.

Keeping the above relation in mind, we apply Lemma 32.7 to $m^2 = a_2^2 + 3b_2^2$ and $m = r_0^2 + 3s_0^2$, and find that there exist coprime integers a_1 and b_1 with $m = \frac{m^2}{m} = a_1^2 + 3b_1^2$ such that

$$a_2 = r_0 a_1 - 3s_0 b_1, \qquad b_2 = r_0 b_1 + s_0 a_1,$$

together with $m \mid (a_1 b_2 - a_2 b_1)$. Combining this with $m \mid (r_0 b_2 - s_0 a_2)$ gives $m \mid (r_0 b_1 - s_0 a_1)$.

Repeating this process once more with Lemma 32.7 applied to $m = a_1^2 + 3b_1^2$ and $m = r_0^2 + 3s_0^2$, we get coprime integers a_0 and b_0 with $1 = \frac{m}{m} = a_0^2 + 3b_0^2$ such that

$$a_1 = r_0 a_0 - 3s_0 b_0, \qquad b_1 = r_0 b_0 + s_0 a_0.$$

However, since $1 = a_0^2 + 3b_0^2$, the only possibilities are $a_0 = \pm 1$ and $b_0 = 0$.

Now we may recover the pairs (a_1, b_1), (a_2, b_2) and (a, b):

$$\begin{cases} a_1 = a_0 r_0, \\ b_1 = a_0 s_0, \end{cases} \qquad \begin{cases} a_2 = a_0(r_0^2 - 3s_0^2), \\ b_2 = 2a_0 r_0 s_0, \end{cases} \qquad \begin{cases} a = a_0 r_0 (r_0 + 3s_0)(r_0 - 3s_0), \\ b = 3a_0 s_0 (r_0 + s_0)(r_0 - s_0). \end{cases}$$

Finally, if $a_0 = 1$, we choose $(r, s) = (r_0, s_0)$, while if $a_0 = -1$, we choose $(r, s) = (-r_0, -s_0)$. ■

Finally, we recover Euler's proof of the cubic case of Fermat's Last Theorem, which shall be stated in a slightly stronger form.

Theorem 32.10 Let X, Y, Z be nonzero integers with $X+Y+Z=0$. Then their product XYZ cannot be a cube.

Proof. Suppose on the contrary that there exist nonzero integers X, Y, Z with $X+Y+Z=0$ such that XYZ is a cube. In particular, we may choose such X, Y, Z with $|XYZ|>0$ minimal.

Note that for this choice, X, Y, Z must be pairwise coprime. This is because if there is some $d>1$ dividing two of them, then it also divides the remaining number. Now we still have that $\frac{X}{d}+\frac{Y}{d}+\frac{Z}{d}$ equals zero and that $\frac{X}{d}\frac{Y}{d}\frac{Z}{d}=\frac{XYZ}{d^3}$ is a cube. Hence, the numbers $\frac{X}{d}, \frac{Y}{d}, \frac{Z}{d}$ provide a counterexample with a smaller absolute product, thereby violating the minimality of $|XYZ|$. From the fact that the product of the pairwise coprime integers X, Y, Z is a cube, we conclude that X, Y, Z themselves are cubes. Write $X=x^3$, $Y=y^3$ and $Z=z^3$. In particular, x, y, z are pairwise coprime, so that exactly one of them is even as we also have $x^3+y^3+z^3=X+Y+Z=0$. Without loss of generality, we assume that x and y are odd and z is even.

Let $x+y=2a$ and $x-y=2b$ so that $x=a+b$ and $y=a-b$. Since x is odd, a and b have different parities. Also, if there is a prime dividing both a and b, then it also divides both x and y, thereby violating the fact that x and y are coprime. Hence, a and b are coprime. Finally, $a, b \neq 0$, for if this is the case, then $x=\pm y$ and hence the only possibilities are $x=\pm 1$ and $y=\pm 1$ as x and y are coprime. But in these cases, we cannot find a nonzero integer z such that $x^3+y^3+z^3=0$.

Note that

$$\begin{aligned} z^3 &= -(x^3+y^3) = -(x+y)(x^2-xy+y^2) \\ &= -(x+y)\left(\frac{(x+y)^2}{4}+\frac{3(x-y)^2}{4}\right) \\ &= -2a(a^2+3b^2). \end{aligned}$$

Since $a, b \neq 0$, we have $a^2+3b^2 \geq 4$ and hence $|-2a|<|z^3|$. Also, as a and b have different parities, a^2+3b^2 is odd. Further, from the fact that a and b are coprime, we see that the greatest common divisor of $-2a$ and a^2+3b^2 is either 1 or 3.

For the first case, we further note that $-2a$ and a^2+3b^2 are nonzero cubes. Applying Corollary 32.9 to the latter gives coprime integers r and s such that $a=r(r+3s)(r-3s)$. Consider the numbers $-2r$, $r+3s$ and $r-3s$, whose sum is zero. It turns out that their product is $-2r(r+3s)(r-3s)=-2a$, which is a nonzero cube as argued earlier, implying that the three numbers are nonzero. Now we have $0<|-2a|<|z^3| \leq |x^3y^3z^3|=|XYZ|$, i.e. the numbers $-2r$, $r+3s$ and $r-3s$ give a smaller absolute product, which is impossible.

For the second case, we have $3 \mid (-2a)$ and $3 \mid (a^2+3b^2)$. Thus, $3 \mid \big(-2a(a^2+3b^2)\big)=z^3$ so that $3 \mid z$, which further gives that $27 \mid z^3=\big(-2a(a^2+3b^2)\big)$. On the other hand, $3 \mid (-2a)$ implies that $3 \mid a$. Since a and b are coprime, we have $3 \nmid b$ so that $9 \nmid (a^2+3b^2)$. It follows that $9 \mid (-2a)$. Further, $-\frac{2a}{9}$ and $\frac{a^2+3b^2}{3}$ are coprime. Meanwhile,

$$\left(\frac{z}{3}\right)^3=-\frac{2a}{9}\cdot\frac{a^2+3b^2}{3},$$

so that $-\frac{2a}{9}$ and $\frac{a^2+3b^2}{3}$ are nonzero cubes. Applying Corollary 32.9 to $\frac{a^2+3b^2}{3}=b^2+3(\frac{a}{3})^2$ gives coprime integers r and s such that $\frac{a}{3}=3r(r+s)(r-s)$. Consider the numbers $-2r$, $r+s$ and $r-s$, whose sum is zero. It turns out that their product is $-2r(r+s)(r-s)=-\frac{2a}{9}$, which is a nonzero cube, implying that the three numbers are nonzero. Now we have $0<|-\frac{2a}{9}|<|z^3| \leq |x^3y^3z^3|=|XYZ|$. Thus, the numbers $-2r$, $r+s$ and $r-s$ also give a smaller absolute product, leading to a contradiction. ■

Bibliography

[1] M. Aigner and G. M. Ziegler, *Proofs from THE BOOK. Sixth Edition*, Springer, Berlin, 2018.

[2] G. E. Andrews, *The Theory of Partitions*, Reprint of the 1976 original, Cambridge University Press, Cambridge, 1998.

[3] T. M. Apostol, *Introduction to Analytic Number Theory*, Springer-Verlag, New York-Heidelberg, 1976.

[4] A. Baker, *Transcendental Number Theory*, Reprint of the 1975 original, Cambridge University Press, Cambridge, 2022.

[5] B. C. Berndt, *Number Theory in the Spirit of Ramanujan*, American Mathematical Society, Providence, RI, 2006.

[6] H. Davenport, *The Higher Arithmetic. An Introduction to the Theory of Numbers. Eighth Edition*, Cambridge University Press, Cambridge, 2008.

[7] L. E. Dickson, *History of the Theory of Numbers. Vols. I–III*, Chelsea Publishing Co., New York, 1966.

[8] G. Gasper and M. Rahman, *Basic Hypergeometric Series. Second Edition*, Cambridge University Press, Cambridge, 2004.

[9] C. F. Gauss, *Disquisitiones Arithmeticae*, Translated by A. A. Clarke, Springer-Verlag, New York, 1986.

[10] R. K. Guy, *Unsolved Problems in Number Theory. Third Edition*, Springer-Verlag, New York, 2004.

[11] H. Halberstam and H.-E. Richert, *Sieve Methods*, No. 4. Academic Press [Harcourt Brace Jovanovich, Publishers], London-New York, 1974.

[12] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers. Sixth Edition*, Oxford University Press, Oxford, 2008.

[13] M. D. Hirschhorn, *The Power of q. A Personal Journey*, Springer, Cham, 2017.

[14] E. Landau, *Elementary Number Theory*, Translated by J. E. Goodman, Chelsea Publishing Co., New York, 1958.

[15] S. Lang, *Algebraic Number Theory. Second Edition*, Springer-Verlag, New York, 1994.

[16] D. A. Marcus, *Number Fields. Second Edition*, Springer, Cham, 2018.

[17] H. L. Montgomery and R. C. Vaughan, *Multiplicative Number Theory. I. Classical Theory*, Cambridge University Press, Cambridge, 2007.

[18] L. J. Mordell, *Diophantine Equations*, Academic Press, London-New York, 1969.

[19] P. Ribenboim, *Fermat's Last Theorem for Amateurs*, Springer-Verlag, New York, 1999.

[20] W. M. Schmidt, *Diophantine Approximations and Diophantine Equations*, Springer-Verlag, Berlin, 1991.

[21] R. C. Vaughan, *The Hardy–Littlewood Method. Second Edition*, Cambridge University Press, Cambridge, 1997.

[22] A. Weil, *Number Theory. An Approach Through History from Hammurapi to Legendre*, Reprint of the 1984 edition, Birkhäuser Boston, Inc., Boston, MA, 2007.

[23] H. S. Wilf, *Generatingfunctionology. Third Edition*, A K Peters, Ltd., Wellesley, MA, 2006.