

where  $C$  is the *Fresnel integral*

$$C = \int_{-\infty}^{\infty} e^{2\pi i t^2} dt.$$

(This is an important example of an infinite integral which converges, although the integrand does not tend to zero.) From (\*) we now obtain the formula for  $G(m, n)$  in the statement of the proposition. To determine the value of the constant  $C$ , take  $m = 1$ ,  $n = 3$ . We obtain  $i\sqrt{3} = \sqrt{3}C(1 + i)$ , which simplifies to  $C = (1 + i)/2$ .  $\square$

From Proposition 10 with  $m = 1$  we obtain

$$G(1, n) = \sum_{v=0}^{n-1} e^{2\pi i v^2/n} = \begin{cases} (1+i)\sqrt{n} & \text{if } n \equiv 0 \pmod{4}, \\ \sqrt{n} & \text{if } n \equiv 1 \pmod{4}, \\ 0 & \text{if } n \equiv 2 \pmod{4}, \\ i\sqrt{n} & \text{if } n \equiv 3 \pmod{4}. \end{cases}$$

If  $m$  and  $n$  are both odd, it follows that

$$\begin{aligned} G(1, mn) &= G(1, m)G(1, n) & \text{if either } m \equiv 1 \text{ or } n \equiv 1 \pmod{4}, \\ &= -G(1, m)G(1, n) & \text{if } m \equiv n \equiv 3 \pmod{4}; \end{aligned}$$

i.e.

$$G(1, mn) = (-1)^{(m-1)(n-1)/4} G(1, m)G(1, n).$$

If, in addition,  $m$  and  $n$  are relatively prime, then  $G(m, n)G(n, m) = G(1, mn)$ , by Proposition 9. Hence, if the integers  $m, n$  are odd, positive and relatively prime, then

$$G(m, n)G(n, m) = (-1)^{(m-1)(n-1)/4} G(1, m)G(1, n).$$

For any odd, positive relatively prime integers  $m, n$ , put

$$\rho(m, n) = G(m, n)/G(1, n).$$

Then

$$\begin{aligned} \rho(1, n) &= 1, \\ \rho(m, n) &= \rho(m', n) & \text{if } m \equiv m' \pmod{n}, \\ \rho(m, n)\rho(n, m) &= (-1)^{(m-1)(n-1)/4}. \end{aligned}$$

We claim that  $\rho(m, n)$  is just the Jacobi symbol  $(m/n)$ . This is evident if  $m = 1$  and, by Proposition 2(i), if  $\rho(m, n) = (m/n)$ , then also  $\rho(n, m) = (n/m)$ .

Hence if the claim is not true for all  $m, n$ , there is a pair  $m, n$  with  $1 < m < n$  such that

$$\rho(m, n) \neq (m/n),$$

but  $\rho(\mu, v) = (\mu/v)$  for all odd, positive relatively prime integers  $\mu, v$  with  $\mu < m$ . We can write  $n = km + r$  for some positive integers  $k, r$  with  $r < m$ .

Then

$$\rho(n, m) = \rho(r, m) = (r/m) = (n/m).$$

Since  $\rho(m, n) \neq (m/n)$ , this yields a contradiction. Thus, if  $n$  is an odd positive integer,

$$G(m, n) = (m/n)G(1, n)$$

for any odd positive integer  $m$  relatively prime to  $n$ .

In fact this relation holds also if  $m$  is negative, since

$$\overline{G(1, n)} = (-1)^{(n-1)/2} G(1, n) \quad \text{and} \quad G(-m, n) = \overline{G(m, n)}.$$

(It may be shown that the relation holds also if  $m$  is even.) As we have already obtained an explicit formula for  $G(1, n)$ , we now have also an explicit evaluation of  $G(m, n)$ .

## 2 Quadratic Fields

Let  $\zeta$  be a complex number which is not rational, but whose square is rational. Since  $\zeta \notin \mathbb{Q}$ , a complex number  $\alpha$  has at most one representation of the form  $\alpha = r + s\zeta$ , where  $r, s \in \mathbb{Q}$ . Let  $\mathbb{Q}(\zeta)$  denote the set of all complex numbers  $\alpha$  which have a representation of this form. Then  $\mathbb{Q}(\zeta)$  is a *field*, since it is closed under subtraction and multiplication and since, if  $r$  and  $s$  are not both zero,

$$(r + s\zeta)^{-1} = (r - s\zeta)/(r^2 - s^2\zeta^2).$$

Evidently  $\mathbb{Q}(\zeta) = \mathbb{Q}(t\zeta)$  for any nonzero rational number  $t$ . Conversely, if  $\mathbb{Q}(\zeta) = \mathbb{Q}(\zeta^*)$ , then  $\zeta^* = t\zeta$  for some nonzero rational number  $t$ . For  $\zeta^* = r + s\zeta$ , where  $r, s \in \mathbb{Q}$  and  $s \neq 0$ , and hence

$$r^2 = \zeta^{*2} - 2s\zeta\zeta^* + s^2\zeta^2.$$

Thus  $\zeta\zeta^*$  is rational, and so is  $\zeta\zeta^*/\zeta^2 = \zeta^*/\zeta$ .

It follows that without loss of generality we may assume that  $\zeta^2 = d$  is a square-free integer. Then  $dt^2 \in \mathbb{Z}$  for some  $t \in \mathbb{Q}$  implies  $t \in \mathbb{Z}$ . If  $\zeta^{*2} = d^*$  is also a square-free integer, then  $\mathbb{Q}(\zeta) = \mathbb{Q}(\zeta^*)$  if and only if  $d = d^*$  and  $\zeta^* = \pm\zeta$ .

The *quadratic field*  $\mathbb{Q}(\sqrt{d})$  is said to be *real* if  $d > 0$  and *imaginary* if  $d < 0$ . We define the *conjugate* of an element  $\alpha = r + s\sqrt{d}$  of the quadratic field  $\mathbb{Q}(\sqrt{d})$  to be the element  $\alpha' = r - s\sqrt{d}$ . It is easily verified that

$$(\alpha + \beta)' = \alpha' + \beta', \quad (\alpha\beta)' = \alpha'\beta'.$$

Since the map  $\sigma : \alpha \rightarrow \alpha'$  is also bijective, it is an *automorphism* of the field  $\mathbb{Q}(\sqrt{d})$ . Since  $\alpha' = \alpha$  if and only if  $s = 0$ , the rational field  $\mathbb{Q}$  is the fixed point set of  $\sigma$ . Since  $(\alpha')' = \alpha$ , the automorphism  $\sigma$  is an 'involution'.

We define the *norm* of an element  $\alpha = r + s\sqrt{d}$  of the quadratic field  $\mathbb{Q}(\sqrt{d})$  to be the rational number

$$N(\alpha) = \alpha\alpha' = r^2 - ds^2.$$

Evidently  $N(\alpha) = N(\alpha')$ , and  $N(\alpha) = 0$  if and only if  $\alpha = 0$ . From the relation  $(\alpha\beta)' = \alpha'\beta'$  we obtain

$$N(\alpha\beta) = N(\alpha)N(\beta).$$

An element  $\alpha$  of the quadratic field  $\mathbb{Q}(\sqrt{d})$  is said to be an *integer* of this field if it is a root of a quadratic polynomial  $t^2 + at + b$  with coefficients  $a, b \in \mathbb{Z}$ . (Equivalently, the integers of  $\mathbb{Q}(\sqrt{d})$  are the elements which are *algebraic integers*.)

It follows from Proposition II.16 that  $\alpha \in \mathbb{Q}(\sqrt{d})$  is an integer of the field  $\mathbb{Q}(\sqrt{d})$  if and only if  $\alpha \in \mathbb{Z}$ . Suppose now that  $\alpha = r + s\sqrt{d}$ , where  $r, s \in \mathbb{Q}$  and  $s \neq 0$ . Then  $\alpha$  is a root of the quadratic polynomial

$$f(x) = (x - \alpha)(x - \alpha') = x^2 - 2rx + r^2 - ds^2.$$

Moreover, this is the unique monic quadratic polynomial with rational coefficients which has  $\alpha$  as a root.

Consequently, if  $\alpha$  is an integer of  $\mathbb{Q}(\sqrt{d})$ , then so also is its conjugate  $\alpha'$  and its norm  $N(\alpha) = r^2 - ds^2$  is an ordinary integer.

**Proposition 11** *Let  $d$  be a square-free integer and define  $\omega$  by*

$$\begin{aligned}\omega &= \sqrt{d} && \text{if } d \equiv 2 \text{ or } 3 \pmod{4}, \\ &= (\sqrt{d} - 1)/2 && \text{if } d \equiv 1 \pmod{4}.\end{aligned}$$

*Then  $\alpha$  is an integer of the quadratic field  $\mathbb{Q}(\sqrt{d})$  if and only if  $\alpha = a + b\omega$  for some  $a, b \in \mathbb{Z}$ .*

*Proof* Suppose  $\alpha = r + s\sqrt{d}$ , where  $r, s \in \mathbb{Q}$ . As we have seen, if  $s = 0$  then  $\alpha$  is an integer of  $\mathbb{Q}(\sqrt{d})$  if and only if  $r \in \mathbb{Z}$ . If  $s \neq 0$ , then  $\alpha$  is an integer of  $\mathbb{Q}(\sqrt{d})$  if and only if  $a = 2r$  and  $b = r^2 - ds^2$  are ordinary integers. If  $a$  is even, i.e. if  $r \in \mathbb{Z}$ , then  $b \in \mathbb{Z}$  if and only if  $ds^2 \in \mathbb{Z}$  and hence, since  $d$  is square-free, if and only if  $s \in \mathbb{Z}$ . If  $a$  is odd, then  $a^2 \equiv 1 \pmod{4}$  and hence  $b \in \mathbb{Z}$  if and only if  $4ds^2 \equiv 1 \pmod{4}$ . Since  $d$  is square-free, this implies that  $2s \in \mathbb{Z}$ ,  $s \notin \mathbb{Z}$ . Hence  $2s$  is odd and  $d \equiv 1 \pmod{4}$ . Conversely, if  $2r$  and  $2s$  are odd integers and  $d \equiv 1 \pmod{4}$ , then  $r^2 - ds^2 \in \mathbb{Z}$ . The result follows.  $\square$

Since  $\omega^2 = -\omega + (d - 1)/4$  in the case  $d \equiv 1 \pmod{4}$ , it follows directly from Proposition 11 that the set  $\mathcal{O}_d$  of all integers of the field  $\mathbb{Q}(\sqrt{d})$  is closed under subtraction and multiplication and consequently is a ring. In fact  $\mathcal{O}_d$  is an integral domain, since  $\mathcal{O}_d \subseteq \mathbb{Q}(\sqrt{d})$ .

For example,  $\mathcal{O}_{-1} = \mathcal{G}$  is the ring of Gaussian integers  $a + bi$ , where  $a, b \in \mathbb{Z}$ . They form a square ‘lattice’ in the complex plane. Similarly  $\mathcal{O}_{-3} = \mathcal{E}$  is the ring of all complex numbers  $a + b\rho$ , where  $a, b \in \mathbb{Z}$  and  $\rho = (i\sqrt{3} - 1)/2$  is a cube root of unity. These *Eisenstein integers* were studied by Eisenstein (1844). They form a hexagonal ‘lattice’ in the complex plane.

We have already seen in §6 of Chapter II that the ring  $\mathcal{G}$  of Gaussian integers is a Euclidean domain, with  $\delta(\alpha) = N(\alpha)$ . We now show that the ring  $\mathcal{E}$  of Eisenstein integers is also a Euclidean domain, with  $\delta(\alpha) = N(\alpha)$ . If  $\alpha, \beta \in \mathcal{E}$  and  $\alpha \neq 0$ , then

$$\beta\alpha^{-1} = \beta\alpha'/\alpha\alpha' = r + sp,$$

where  $r, s \in \mathbb{Q}$ . Choose  $a, b \in \mathbb{Z}$  so that

$$|r - a| \leq 1/2, \quad |s - b| \leq 1/2.$$

If  $\kappa = a + b\rho$ , then  $\kappa \in \mathcal{E}$  and

$$\begin{aligned} N(\beta\alpha^{-1} - \kappa) &= \{r - a - (s - b)/2\}^2 + 3\{(s - b)/2\}^2 \\ &\leq (3/4)^2 + 3(1/4)^2 = 3/4 < 1. \end{aligned}$$

Thus  $\beta - \kappa\alpha \in \mathcal{E}$  and  $N(\beta - \kappa\alpha) < N(\alpha)$ .

Since  $\mathcal{G}$  and  $\mathcal{E}$  are Euclidean domains, the divisibility theory of Chapter II is valid for them. As an application, we prove

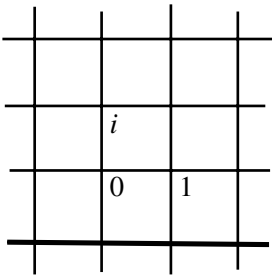
**Proposition 12** *The equation  $x^3 + y^3 = z^3$  has no solutions in nonzero integers.*

*Proof* Assume on the contrary that such a solution exists and choose one for which  $|xyz|$  is a minimum. Then  $(x, y) = (x, z) = (y, z) = 1$ . If 3 did not divide  $xyz$ , then  $x^3, y^3$  and  $z^3$  would be congruent to  $\pm 1 \pmod 9$ , which contradicts  $x^3 + y^3 = z^3$ . So, without loss of generality, we may assume that  $3|z$ . Then  $x^3 + y^3 \equiv 0 \pmod 3$  and, again without loss of generality, we may assume that  $x \equiv 1 \pmod 3, y \equiv -1 \pmod 3$ . This implies that

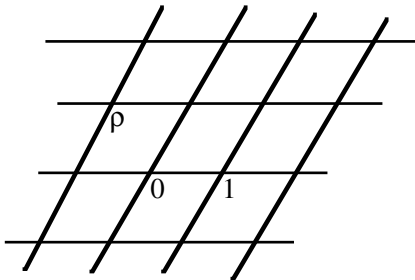
$$x^2 - xy + y^2 \equiv 3 \pmod 9.$$

If  $x + y$  and  $x^2 - xy + y^2$  have a common prime divisor  $p$ , then  $p$  divides  $3xy$ , since  $3xy = (x + y)^2 - (x^2 - xy + y^2)$ , and this implies  $p = 3$ , since  $(x, y) = 1$ . Since

$$(x + y)(x^2 - xy + y^2) = x^3 + y^3 = z^3 \equiv 0 \pmod{27},$$



$\mathcal{G} = \mathcal{O}_{-1}$ : Gaussian integers



$\mathcal{E} = \mathcal{O}_{-3}$ : Eisenstein integers

Fig. 1. Gaussian and Eisenstein integers.

it follows that

$$\begin{aligned}x + y &= 9a^3, \\x^2 - xy + y^2 &= 3b^3,\end{aligned}$$

where  $a, b \in \mathbb{Z}$  and  $3 \nmid b$ .

We now shift operations to the Euclidean domain  $\mathcal{E}$  of Eisenstein integers. We have

$$x^2 - xy + y^2 = (x + y\rho)(x + y\rho^2),$$

where  $\rho = (i\sqrt{3} - 1)/2$  is a cube root of unity. Put  $\lambda = 1 - \rho$ , so that  $(1 + \rho)\lambda^2 = 3$ . Then  $\lambda$  is a common divisor of  $x + y\rho$  and  $x + y\rho^2$ , since

$$\begin{aligned}x + y\rho &= x + y - y\lambda, \\x + y\rho^2 &= x - 2y + y\lambda\end{aligned}$$

and  $x + y \equiv 0 \equiv x - 2y \pmod{3}$ . In fact  $\lambda$  is the greatest common divisor of  $x + y\rho$  and  $x + y\rho^2$  since, for all  $m, n \in \mathbb{Z}$ ,

$$(m + n + n\rho)(x + y\rho^2) - (n + mp + n\rho)(x + y\rho) = (mx + ny)\lambda$$

and we can choose  $m, n$  so that  $mx + ny = 1$ . Since  $\lambda^2 = -3\rho$  and since  $\rho$  is a unit, from  $(x + y\rho)(x + y\rho^2) = 3b^3$  and the unique factorization of  $b$  in  $\mathcal{E}$ , we now obtain

$$x + y\rho = \varepsilon\lambda(c + d\rho)^3,$$

where  $c, d \in \mathbb{Z}$  and  $\varepsilon$  is a unit. From

$$(x + y\rho)/\lambda = x - \lambda(x + y)/3 = x - 3a^3\lambda$$

and

$$(c + d\rho)^3 = c^3 - 3cd^2 + d^3 + 3cd(c - d)\rho,$$

by reducing mod 3 we get

$$\varepsilon(c^3 + d^3) \equiv 1 \pmod{3}.$$

Since the units in  $\mathcal{E}$  are  $\pm 1, \pm\rho, \pm\rho^2$  (by the following Proposition 13), this implies  $\varepsilon = \pm 1$ . In fact we may suppose  $\varepsilon = 1$ , by changing the signs of  $c$  and  $d$ . Equating coefficients of  $\rho$ , we now get

$$a^3 = cd(c - d).$$

But  $(c, d) = 1$ , since  $(x, y) = 1$ , and hence also  $(c, c - d) = (d, c - d) = 1$ . It follows that  $c = z_1^3, d = y_1^3, c - d = x_1^3$  for some  $x_1, y_1, z_1 \in \mathbb{Z}$ . Thus  $x_1^3 + y_1^3 = z_1^3$  and

$$|x_1 y_1 z_1| = |a| = |z/3b| < |xyz|.$$

But this contradicts the definition of  $x, y, z$ . □

The proof of Proposition 12 illustrates how problems involving ordinary integers may be better understood by viewing them as integers in a larger field of algebraic numbers.

We now return to the study of an arbitrary quadratic field  $\mathbb{Q}(\sqrt{d})$ , where  $d$  is a square-free integer. For convenience of writing we put  $J = \mathcal{O}_d$ . As in Chapter II, we say that  $\varepsilon \in J$  is a *unit* if there exists  $\eta \in J$  such that  $\varepsilon\eta = 1$ . For example, 1 and  $-1$  are units. The set  $U$  of all units is evidently an abelian group under multiplication. Moreover, if  $\varepsilon \in U$ , then also  $\varepsilon' \in U$ .

If  $\varepsilon$  is a unit, then  $N(\varepsilon) = \pm 1$ , since  $\varepsilon\eta = 1$  implies  $N(\varepsilon)N(\eta) = 1$ . Conversely, if  $\varepsilon \in J$  and  $N(\varepsilon) = \pm 1$ , then  $\varepsilon$  is a unit, since  $N(\varepsilon) = \varepsilon\varepsilon'$  and  $\varepsilon' \in J$ . (Note, however, that  $N(\alpha) = \pm 1$  does not imply  $\alpha \in J$ . For example, in  $\mathbb{Q}(\sqrt{-1})$ ,  $\alpha = (3+4i)/5 \notin \mathcal{O}$ , although  $N(\alpha) = 1$ .) It follows that, when  $d \equiv 2$  or  $3 \pmod{4}$ ,  $\alpha = a + b\sqrt{d}$  is a unit if and only if  $a, b \in \mathbb{Z}$  and

$$a^2 - db^2 = \pm 1.$$

On the other hand, when  $d \equiv 1 \pmod{4}$ ,  $\alpha = a + b(\sqrt{d} - 1)/2$  is a unit if and only if  $a, b \in \mathbb{Z}$  and

$$(b - 2a)^2 - db^2 = \pm 4.$$

But if  $b, c \in \mathbb{Z}$  and  $c^2 - db^2 = \pm 4$ , then  $c^2 \equiv b^2 \pmod{4}$  and hence  $c \equiv b \pmod{2}$ .

Consequently, the units of  $J$  are determined by the solutions of the Diophantine equations  $x^2 - dy^2 = \pm 4$  or  $x^2 - dy^2 = \pm 1$ , according as  $d \equiv 1$  or  $d \not\equiv 1 \pmod{4}$ . This makes it possible to determine all units, as we now show.

**Proposition 13** *The units of  $\mathcal{O}_{-1}$  are  $\pm 1, \pm i$  and the units of  $\mathcal{O}_{-3}$  are  $\pm 1, (\pm 1 \pm i\sqrt{3})/2$ . For every other square-free integer  $d < 0$ , the only units of  $\mathcal{O}_d$  are  $\pm 1$ .*

*For each square-free integer  $d > 0$ , there exists a unit  $\varepsilon_0 > 1$  such that all units of  $\mathcal{O}_d$  are given by  $\pm \varepsilon_0^n$  ( $n \in \mathbb{Z}$ ).*

*Proof* Suppose first that  $d < 0$ . Then only the Diophantine equations with the  $+$  signs need to be considered. If  $d < -4$ , the only solutions of  $x^2 - dy^2 = 4$  are  $y = 0, x = \pm 2$ . If  $d < -4$  or if  $d = -2$ , the only solutions of  $x^2 - dy^2 = 1$  are  $y = 0, x = \pm 1$ . In these cases the only units are  $\pm 1$ . (The group  $U$  is a cyclic group of order 2, with  $-1$  as generator.) If  $d = -3$ , the only solutions of  $x^2 - dy^2 = 4$  are  $y = 0, x = \pm 2$  and  $y = \pm 1, x = \pm 1$ . Hence the units are  $\pm 1, \pm \rho, \pm \rho^2$ , where  $\rho = (i\sqrt{3} - 1)/2$ . (The group  $U$  is a cyclic group of order 6, with  $-\rho$  as generator.) If  $d = -1$ , the only solutions of  $x^2 + y^2 = 1$  are  $y = 0, x = \pm 1$  and  $y = \pm 1, x = 0$ . Hence the units are  $\pm 1, \pm i$ . (The group  $U$  is a cyclic group of order 4, with  $i$  as generator.)

Suppose next that  $d > 0$ . With the aid of continued fractions it will be shown in §4 of Chapter IV that the equation  $x^2 - dy^2 = 1$  always has a solution in positive integers and, by doubling them, so also does the equation  $x^2 - dy^2 = 4$ . Hence there always exists a unit  $\varepsilon > 1$ . For any unit  $\varepsilon > 1$  we have  $\varepsilon > \pm \varepsilon'$ , since  $\varepsilon' = \varepsilon^{-1}$  or  $-\varepsilon^{-1}$ . If  $\varepsilon = a + b\omega$ , where  $\omega$  is defined as in Proposition 11 and  $a, b \in \mathbb{Z}$ , then  $\varepsilon' = a - b\omega$  or  $a - b - b\omega$ , according as  $d \not\equiv 1$  or  $d \equiv 1 \pmod{4}$ . Since  $\omega$  is positive,  $\varepsilon > \varepsilon'$  yields  $b > 0$  and  $\varepsilon > -\varepsilon'$  then yields  $a > 0$ . Thus every unit  $\varepsilon > 1$  has the form  $a + b\omega$ , where

$a, b \in \mathbb{N}$ . Consequently there is a least unit  $\varepsilon_0 > 1$ . Then, for any unit  $\varepsilon > 1$ , there is a positive integer  $n$  such that  $\varepsilon_0^n \leq \varepsilon < \varepsilon_0^{n+1}$ . Since  $\varepsilon\varepsilon_0^{-n}$  is a unit and  $1 \leq \varepsilon\varepsilon_0^{-n} < \varepsilon_0$ , we must actually have  $\varepsilon = \varepsilon_0^n$ . (The group  $U$  is the direct product of the cyclic group of order 2 generated by  $-1$  and the infinite cyclic group generated by  $\varepsilon_0$ .)  $\square$

As an example, take  $d = 2$ . Then  $\varepsilon_0 = 1 + \sqrt{2}$  is a unit. Since  $\varepsilon_0 > 1$  and all units greater than 1 have the form  $a + b\sqrt{2}$  with  $a, b \in \mathbb{N}$ , it follows that all units are given by  $\pm\varepsilon_0^n$  ( $n \in \mathbb{Z}$ ).

Having determined the units, we now consider more generally the theory of divisibility in the integral domain  $J$ . If  $\alpha, \beta \in J$  and  $\beta$  is a proper divisor of  $\alpha$ , then  $N(\beta)$  is a proper divisor in  $\mathbb{Z}$  of  $N(\alpha)$  and hence  $|N(\beta)| < |N(\alpha)|$ . Consequently the chain condition (#) of Chapter II is satisfied. It follows that any element of  $J$  which is neither zero nor a unit is a product of finitely many irreducibles. Thus it only remains to determine the irreducibles. However, this is not such a simple matter, as the following examples indicate.

The ring  $\mathcal{G}$  of Gaussian integers is a Euclidean domain. However, an ordinary prime  $p$  may or may not be irreducible in  $\mathcal{G}$ . For example,  $2 = (1 + i)(1 - i)$  and neither factor is one of the units  $\pm 1, \pm i$ . On the other hand, 3 has no proper divisor  $\alpha = a + bi$  which is not a unit, since  $N(3) = 9$  and  $N(\alpha) = a^2 + b^2 = \pm 3$  has no solutions in integers  $a, b$ .

Again, consider the ring  $\mathcal{O}_{-5}$  of integers of the field  $\mathbb{Q}(\sqrt{-5})$ . An element  $\alpha = a + b\sqrt{-5}$  of  $\mathcal{O}_{-5}$  cannot have norm  $N(\alpha) = a^2 + 5b^2$  equal to  $\pm 2$  or  $\pm 3$ , since the square of any ordinary integer is congruent to 0, 1 or 4 mod 5. It follows that, in the factorizations

$$6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5}),$$

all four factors are irreducible and the factorizations are essentially distinct, since  $N(2) = 4$ ,  $N(3) = 9$  and  $N(1 \pm \sqrt{-5}) = 6$ . Thus 2 is not a ‘prime’ in  $\mathcal{O}_{-5}$  and the ‘fundamental theorem of arithmetic’ does not hold.

It was shown by Kummer and Dedekind in the 19th century that uniqueness of factorization could be restored by considering ideals instead of elements. Any nonzero proper ideal of  $\mathcal{O}_d$  can be represented as a product of finitely many prime ideals and the representation is unique except for the order of the factors. This result will now be established.

A nonempty subset  $A$  of a commutative ring  $R$  is an *ideal* if  $a, b \in A$  and  $x, y \in R$  imply  $ax + by \in A$ . For example,  $R$  and  $\{0\}$  are ideals. If  $a_1, \dots, a_m \in R$ , then the set  $(a_1, \dots, a_m)$  of all elements  $a_1x_1 + \dots + a_mx_m$  with  $x_j \in R$  ( $1 \leq j \leq m$ ) is an ideal, the ideal *generated* by  $a_1, \dots, a_m$ . An ideal generated by a single element is a *principal ideal*.

If  $A$  and  $B$  are ideals in  $R$ , then the set  $AB$  of all finite sums  $a_1b_1 + \dots + a_nb_n$  with  $a_j \in A$  and  $b_j \in B$  ( $1 \leq j \leq n$ ;  $n \in \mathbb{N}$ ) is also an ideal, the *product* of  $A$  and  $B$ . For any ideals  $A, B, C$  we have

$$AB = BA, (AB)C = A(BC),$$

since multiplication in  $R$  is commutative and associative.

An ideal  $A \neq \{0\}$  is said to be *divisible* by an ideal  $B$ , and  $B$  is said to be a *factor* of  $A$ , if there exists an ideal  $C$  such that  $A = BC$ . For example,  $A$  is divisible by itself and by  $R$ , since  $A = AR$ . Thus  $R$  is an identity element for multiplication of ideals.

Now take  $R = \mathcal{O}_d$  to be the ring of all integers of the quadratic field  $\mathbb{Q}(\sqrt{d})$ . We will show that in this case much more can be said.

**Proposition 14** *Let  $A \neq \{0\}$  be an ideal in  $\mathcal{O}_d$ . Then there exist  $\beta, \gamma \in A$  such that every  $\alpha \in A$  can be uniquely represented in the form*

$$\alpha = m\beta + n\gamma \quad (m, n \in \mathbb{Z}).$$

*Furthermore, if  $\omega$  is defined as in Proposition 11, we may take  $\beta = a$ ,  $\gamma = b + c\omega$ , where  $a, b, c \in \mathbb{Z}$ ,  $a > 0$ ,  $c > 0$ ,  $c$  divides both  $a$  and  $b$ , and  $ac$  divides  $\gamma\gamma'$ , i.e.*

$$\begin{aligned} b^2 - dc^2 &\equiv 0 \pmod{ac} & \text{if } d \equiv 2 \text{ or } 3 \pmod{4}, \\ b(b - c) - (d - 1)c^2/4 &\equiv 0 \pmod{ac} & \text{if } d \equiv 1 \pmod{4}. \end{aligned}$$

*Proof* Since  $A$  is an ideal, the set  $J$  of all  $z \in \mathbb{Z}$  such that  $y + z\omega \in A$  for some  $y \in \mathbb{Z}$  is an ideal in  $\mathbb{Z}$ . Moreover  $J \neq \{0\}$ , since  $A \neq \{0\}$  and  $\alpha \in A$  implies  $\alpha\omega \in A$ . Since  $\mathbb{Z}$  is a principal ideal domain, it follows that there exists  $c > 0$  such that  $J = \{nc : n \in \mathbb{Z}\}$ . Since  $c \in J$ , there exists  $b \in \mathbb{Z}$  such that  $\gamma := b + c\omega \in A$ .

Moreover  $A$  contains some nonzero  $x \in \mathbb{Z}$ , since  $\alpha \in A$  implies  $\alpha\alpha' \in A$ . Since the set  $I$  of all  $x \in \mathbb{Z} \cap A$  is an ideal in  $\mathbb{Z}$ , there exists  $a > 0$  such that  $I = \{ma : m \in \mathbb{Z}\}$ . For any  $\alpha = y + z\omega \in A$  we have  $z = nc$  for some  $n \in \mathbb{Z}$  and  $\alpha - n\gamma = y - nb = ma$  for some  $m \in \mathbb{Z}$ . Thus  $\alpha = m\beta + n\gamma$  with  $\beta = a$ . The representation is unique, since  $\gamma$  is irrational.

Since  $\beta\omega \in A$ , we have

$$a\omega = ra + s(b + c\omega) \quad \text{for unique } r, s \in \mathbb{Z}.$$

Thus  $a = sc$  and  $ra + sb = 0$ , which together imply  $b = -rc$ . Since  $\gamma\omega \in A$ , we have also

$$(b + c\omega)\omega = ma + n(b + c\omega) \quad \text{for unique } m, n \in \mathbb{Z}.$$

If  $d \equiv 2$  or  $3 \pmod{4}$ , then  $\omega^2 = d$ . In this case  $n = -r$ ,  $cd = ma - rb$  and hence  $dc^2 = mac + b^2$ . If  $d \equiv 1 \pmod{4}$ , then  $\omega^2 = -\omega + (d - 1)/4$ . Hence  $n = -(r + 1)$ ,  $c(d - 1)/4 = ma - rb - b$  and  $(d - 1)c^2/4 = mac + b(b - c)$ .  $\square$

If  $A$  is an ideal in  $\mathcal{O}_d$ , then the set  $A' = \{\alpha' : \alpha \in A\}$  of all conjugates of elements of  $A$  is also an ideal in  $\mathcal{O}_d$ . We call  $A'$  the *conjugate* of  $A$ .

**Proposition 15** *If  $A \neq \{0\}$  is an ideal in  $\mathcal{O}_d$ , then  $AA' = l\mathcal{O}_d$  for some  $l \in \mathbb{N}$ .*

*Proof* Choose  $\beta, \gamma$  so that  $A = \{m\beta + n\gamma : m, n \in \mathbb{Z}\}$ . Then  $AA'$  consists of all integral linear combinations of  $\beta\beta', \beta\gamma', \beta'\gamma$  and  $\gamma\gamma'$ . Furthermore  $r = \beta\beta'$ ,  $s = \beta\gamma' + \beta'\gamma$  and  $t = \gamma\gamma'$  are all in  $\mathbb{Z}$ . If  $l$  is the greatest common divisor of  $r, s$  and  $t$ , then  $l \in AA'$ , by the Bézout identity, and hence  $l\mathcal{O}_d \subseteq AA'$ .

On the other hand,  $\beta\gamma'$  and  $\beta'\gamma$  are roots of the quadratic equation

$$x^2 - sx + rt = 0$$

with integer coefficients  $s = \beta\gamma' + \beta'\gamma$  and  $rt = \beta\beta'\gamma\gamma'$ . It follows that  $\beta\gamma'/l$  and  $\beta'\gamma/l$  are roots of the quadratic equation

$$y^2 - (s/l)y + rt/l^2 = 0,$$

which also has integer coefficients. Since  $\beta\gamma'/l$  and  $\beta'\gamma/l$  are in  $\mathbb{Q}(\sqrt{d})$ , this means that they are in  $\mathcal{O}_d$ . Thus  $\beta\gamma'$  and  $\beta'\gamma$  are in  $l\mathcal{O}_d$ . Since also  $\beta\beta'$  and  $\gamma\gamma'$  are in  $l\mathcal{O}_d$ , it follows that  $AA' \subseteq l\mathcal{O}_d$ .  $\square$

If in the proof of Proposition 15 we choose  $\beta = a$  and  $\gamma = b + c\omega$  as in the statement of Proposition 14, then in the statement of Proposition 15 we will have  $l = ac$ . Since the proof of this when  $d \equiv 1 \pmod{4}$  is similar, we give the proof only for  $d \equiv 2$  or  $3 \pmod{4}$ . In this case  $\omega = \sqrt{d}$  and hence  $r = a^2$ ,  $s = 2ab$ ,  $t = b^2 - dc^2$ . We wish to show that  $ac$  is the greatest common divisor of  $r$ ,  $s$  and  $t$ . Thus if we put

$$a = cu, b = cv, t = acw,$$

then we wish to show that  $u$ ,  $2v$  and  $w$  have greatest common divisor 1. Since  $uw = v^2 - d$  and  $d$  is square-free, a common divisor greater than 1 can only be 2. But if 2 were a common divisor, we would have  $v^2 \equiv d \pmod{4}$ , which is impossible, because  $d \equiv 2$  or  $3 \pmod{4}$ .

We can now show that multiplication of ideals satisfies the cancellation law:

**Proposition 16** *If  $A, B, C$  are ideals in  $\mathcal{O}_d$  with  $A \neq \{0\}$ , then  $AB = AC$  implies  $B = C$ .*

*Proof* By multiplying by the conjugate  $A'$  of  $A$  we obtain  $AA'B = AA'C$  and hence, by Proposition 15,  $lB = lC$  for some positive integer  $l$ . But this implies  $B = C$ .  $\square$

**Proposition 17** *Let  $A$  and  $B$  be nonzero ideals in  $\mathcal{O}_d$ . Then  $A$  is divisible by  $B$  if and only if  $A \subseteq B$ .*

*Proof* If  $A = BC$  for some ideal  $C$ , then  $A \subseteq B$ , by the definition of the product of two ideals.

Conversely, suppose  $A \subseteq B$ . By Proposition 15,  $BB' = l\mathcal{O}_d$  for some positive integer  $l$ . Hence  $AB' \subseteq l\mathcal{O}_d$ . It follows that  $AB' = lC$  for some ideal  $C$ . Thus  $AB' = BB'C$  and so, by Proposition 16,  $A = BC$ .  $\square$

**Corollary 18** *Let  $A$  and  $B$  be nonzero ideals in  $\mathcal{O}_d$ . If  $D$  is the set of all elements  $a + b$ , with  $a \in A$  and  $b \in B$ , then  $D$  is an ideal and is a factor of both  $A$  and  $B$ . Moreover, every common factor of  $A$  and  $B$  is also a factor of  $D$ .*

*Proof* It follows at once from its definition that  $D$  is an ideal. Moreover  $D$  contains both  $A$  and  $B$ , since 0 is an element of any ideal. Evidently also any ideal  $C$  which contains both  $A$  and  $B$  also contains  $D$ . The result now follows from Proposition 17.  $\square$

In the terminology of Chapter II, §1, this shows that any two nonzero ideals in  $\mathcal{O}_d$  have a greatest common divisor.

In a commutative ring  $R$ , an ideal  $A \neq R, \{0\}$  is said to be *irreducible* if its only factors are  $A$  and  $R$ . It is said to be *maximal* if the only ideals containing  $A$  are  $A$  and  $R$ . It is said to be *prime* if, whenever  $A$  divides the product of two ideals, it also divides at least one of the factors.

By Proposition 17, an ideal in  $\mathcal{O}_d$  is irreducible if and only if it is maximal. As we saw in §1 of Chapter II, the existence of greatest common divisors implies that an ideal in  $\mathcal{O}_d$  is irreducible if and only if it is prime. (These equivalences do not hold in all commutative rings, but they do hold for the ring of all algebraic integers in any given algebraic number field, and also for the rings associated with algebraic curves.)

**Proposition 19** *A nonzero ideal  $A$  in  $\mathcal{O}_d$  has only finitely many factors.*

*Proof* Since  $AA' = l\mathcal{O}_d$  for some positive integer  $l$ , any factor  $B$  of  $A$  is also a factor of  $l\mathcal{O}_d$  and so contains  $l$ . Proposition 14 implies, in particular, that  $B$  is generated by two elements, say  $B = (\beta_1, \beta_2)$ . *A fortiori*,  $B = (\beta_1, \beta_2, l)$  and hence, for any  $\gamma_1, \gamma_2 \in \mathcal{O}_d$ , also

$$B = (\beta_1 - l\gamma_1, \beta_2 - l\gamma_2, l).$$

We can choose  $\gamma_1 \in \mathcal{O}_d$  so that in the representation

$$\beta_1 - l\gamma_1 = a_1 + b_1\omega \quad (a_1, b_1 \in \mathbb{Z})$$

we have  $0 \leq a_1, b_1 < l$ . Similarly we can choose  $\gamma_2 \in \mathcal{O}_d$  so that in the representation

$$\beta_2 - l\gamma_2 = a_2 + b_2\omega \quad (a_2, b_2 \in \mathbb{Z})$$

we have  $0 \leq a_2, b_2 < l$ . It follows that there are at most  $l^4$  different possibilities for the ideal  $B$ .  $\square$

**Corollary 20** *There exists no infinite sequence  $\{A_n\}$  of nonzero ideals in  $\mathcal{O}_d$  such that, for every  $n$ ,  $A_{n+1}$  divides  $A_n$  and  $A_{n+1} \neq A_n$ .*

In the terminology of Chapter II, this shows that the set of all nonzero ideals in  $\mathcal{O}_d$  satisfies the chain condition (#). Since also the conclusion of Proposition II.1 holds, the argument in §1 of Chapter II now shows that any nonzero proper ideal in  $\mathcal{O}_d$  is a product of finitely many prime ideals and the representation is unique apart from the order of the factors.

It remains to determine the prime ideals. This is accomplished by the following three propositions.

**Proposition 21** *For each prime ideal  $P$  in  $\mathcal{O}_d$  there is a unique prime number  $p$  such that  $P$  divides  $p\mathcal{O}_d$ . Furthermore, for any prime number  $p$  there is a prime ideal  $P$  in  $\mathcal{O}_d$  such that exactly one of the following alternatives holds:*

- (i)  $p\mathcal{O}_d = PP'$  and  $P \neq P'$ ;
- (ii)  $p\mathcal{O}_d = P = P'$ ;
- (iii)  $p\mathcal{O}_d = P^2$  and  $P = P'$ .

*Proof* If  $P$  is a prime ideal in  $\mathcal{O}_d$ , then  $PP' = l\mathcal{O}_d$  for some positive integer  $l$ . Moreover  $l > 1$ , since  $l \in P$ . If  $l = mn$ , where  $m$  and  $n$  are positive integers greater than 1, then  $P$  divides either  $m\mathcal{O}_d$  or  $n\mathcal{O}_d$ . By repeating the argument it follows that  $P$  divides  $p\mathcal{O}_d$  for some prime divisor  $p$  of  $l$ . The prime number  $p$  is uniquely determined by the prime ideal  $P$  since, by the Bézout identity, if  $P$  contained distinct primes it would also contain their greatest common divisor 1.

Now let  $p$  be any prime number and let the factorisation of  $p\mathcal{O}_d$  into a product of positive powers of distinct prime ideals be

$$p\mathcal{O}_d = P_1^{e_1} \cdots P_s^{e_s}.$$

If we put  $Q_j = P'_j$  ( $1 \leq j \leq s$ ), then also

$$p\mathcal{O}_d = Q_1^{e_1} \cdots Q_s^{e_s}.$$

But  $P_j Q_j = n_j \mathcal{O}_d$  for some integer  $n_j > 1$  and hence

$$p^2 = n_1^{e_1} \cdots n_s^{e_s}.$$

Evidently the only possibilities are

- (i)'  $s = 2, n_1 = n_2 = p, e_1 = e_2 = 1$ ;
- (ii)'  $s = 1, n_1 = p^2, e_1 = 1$ ;
- (iii)'  $s = 1, n_1 = p, e_1 = 2$ .

Since the factorisation is unique apart from order, this yields the three possibilities in the statement of the proposition.  $\square$

Proposition 21 does not tell us which of the three possibilities holds for a given prime  $p$ . For  $p \neq 2$ , the next result gives an answer in terms of the Legendre symbol.

**Proposition 22** *Let  $p$  be an odd prime. Then, in the statement of Proposition 21, (i), (ii), or (iii) holds according as*

$$p \nmid d \text{ and } (d/p) = 1, \quad p \nmid d \text{ and } (d/p) = -1, \quad \text{or } p|d.$$

*Proof* Suppose first that  $p \nmid d$  and that there exists  $a \in \mathbb{Z}$  such that  $a^2 \equiv d \pmod{p}$ . Then  $p \nmid a$  and  $a^2 - d = pb$  for some  $b \in \mathbb{Z}$ . If  $P = (p, a + \sqrt{d})$ , then  $P' = (p, a - \sqrt{d})$  and

$$PP' = p(p, a + \sqrt{d}, a - \sqrt{d}, b).$$

Since  $(p, a + \sqrt{d}, a - \sqrt{d}, b)$  contains  $2a$ , which is relatively prime to  $p$ , it also contains 1. Hence  $PP' = p\mathcal{O}_d$ . Furthermore  $P \neq P'$ , since  $P = P'$  would imply  $2a \in P$  and hence  $1 \in P$ . We do not need to prove that  $P$  is a prime ideal, since what we have already established is incompatible with cases (ii) and (iii) of Proposition 21.

Suppose next that  $p|d$ . Then  $d = pe$  for some  $e \in \mathbb{Z}$  and  $p \nmid e$ , since  $d$  is square-free. If  $P = (p, \sqrt{d})$ , then

$$P^2 = p(p, \sqrt{d}, e) = p\mathcal{O}_d,$$

since  $(p, e) = 1$ . Since we cannot be in cases (i) or (ii) of Proposition 21, we must be in case (iii).

Suppose conversely that either (i) or (iii) of Proposition 21 holds. Then the corresponding prime ideal  $P$  contains  $p$ . Choose  $\beta = a$  and  $\gamma = b + c\omega$  as in Proposition 14, so that

$$P = \{m\beta + n\gamma : m, n \in \mathbb{Z}\}.$$

In the present case we must have  $a = p$ , since  $p \in P$  and  $1 \notin P$ . We must also have  $c = 1$ , since  $PP' = p\mathcal{O}_d$  implies  $ac = p$ . With these values of  $a$  and  $c$  the final condition of Proposition 14 takes the form

$$\begin{aligned} b^2 &\equiv d \pmod{p} && \text{if } d \equiv 2 \text{ or } 3 \pmod{4}, \\ b(b-1) &\equiv (d-1)/4 \pmod{p} && \text{if } d \equiv 1 \pmod{4}. \end{aligned}$$

Thus in the latter case  $(2b-1)^2 \equiv d \pmod{p}$ . In either case if  $p \nmid d$ , then  $(d/p) = 1$ .

This proves that if  $p \nmid d$  and  $(d/p) = -1$ , then we must be in case (ii) of Proposition 21.  $\square$

**Proposition 23** *Let  $p = 2$ . Then, in the statement of Proposition 21, (i), (ii), or (iii) holds according as*

$$d \equiv 1 \pmod{8}, d \equiv 5 \pmod{8}, \text{ or } d \equiv 2, 3 \pmod{4}.$$

*Proof* Since the proof is similar to that of the previous proposition, we will omit some of the detail. Suppose first that  $d \equiv 1 \pmod{8}$ . If  $P = (2, (1 - \sqrt{d})/2)$ , then  $P' = (2, (1 + \sqrt{d})/2)$  and

$$PP' = 2(2, (1 - \sqrt{d})/2, (1 + \sqrt{d})/2, (1 - d)/8).$$

It follows that  $PP' = 2\mathcal{O}_d$  and  $P \neq P'$ .

Suppose next that  $d \equiv 2 \pmod{4}$ . Then  $d = 2e$ , where  $e$  is odd. If  $P = (2, \sqrt{d})$ , then

$$P^2 = 2(2, \sqrt{d}, e) = 2\mathcal{O}_d.$$

Similarly, if  $d \equiv 3 \pmod{4}$  and  $P = (2, 1 + \sqrt{d})$ , then

$$P^2 = 2(2, 1 + \sqrt{d}, (1 + d)/2 + \sqrt{d}) = 2\mathcal{O}_d.$$

Suppose conversely that either (i) or (iii) of Proposition 21 holds. Then the corresponding prime ideal  $P$  contains 2. Choose  $\beta = a$  and  $\gamma = b + c\omega$  as in Proposition 14, so that

$$P = \{m\beta + n\gamma : m, n \in \mathbb{Z}\}.$$

In the present case we must have  $a = 2$ ,  $c = 1$  and

$$b(b-1) \equiv (d-1)/4 \pmod{2} \quad \text{if } d \equiv 1 \pmod{4}.$$

Since  $b(b-1)$  is even, it follows that  $d \not\equiv 5 \pmod{8}$ .

This proves that if  $d \equiv 5 \pmod{8}$ , then we must be in case (ii) of Proposition 21.  $\square$

Proposition 22 uses only Legendre's definition of the Legendre symbol. What does the law of quadratic reciprocity tell us? By Proposition 4, if  $p$  and  $q$  are distinct odd primes and  $d$  an integer not divisible by  $p$  such that  $q \equiv p \pmod{4d}$ , then  $(d/p) = (d/q)$ . Consequently, by Proposition 22, whether (i) or (ii) holds in Proposition 21 depends only on the residue class of  $p \pmod{4d}$ . Thus, for given  $d$ , we need determine the behaviour of only finitely many primes  $p$ .

We mention without proof some further properties of the ring  $\mathcal{O}_d$ . We say that two nonzero ideals  $A, \tilde{A}$  in  $\mathcal{O}_d$  are *equivalent*, and we write  $A \sim \tilde{A}$ , if there exist nonzero principal ideals  $(\alpha), (\tilde{\alpha})$  such that  $(\alpha)A = (\tilde{\alpha})\tilde{A}$ . It is easily verified that this is indeed an equivalence relation. Moreover, if  $A \sim \tilde{A}$  and  $B \sim \tilde{B}$ , then  $AB \sim \tilde{A}\tilde{B}$ . Consequently, if we call an equivalence class of ideals an *ideal class*, we can without ambiguity define the product of two ideal classes. The set of ideal classes acquires in this way the structure of a commutative group, the ideal class containing the conjugate  $A'$  of  $A$  being the inverse of the ideal class containing  $A$ . It may be shown that this *ideal class group* is finite. The order of the group, i.e. the number of different ideal classes, is called the *class number* of the quadratic field  $\mathbb{Q}(\sqrt{d})$  and is traditionally denoted by  $h(d)$ . The ring  $\mathcal{O}_d$  is a principal ideal domain if and only if  $h(d) = 1$ . (It may be shown that  $\mathcal{O}_d$  is a factorial domain only if it is a principal ideal domain.)

The theory of quadratic fields has been extensively generalized. An *algebraic number field*  $K$  is a field containing the field  $\mathbb{Q}$  of rational numbers and of finite dimension as a vector space over  $\mathbb{Q}$ . An *algebraic integer* is a root of a monic polynomial  $x^n + a_1x^{n-1} + \cdots + a_n$  with coefficients  $a_1, \dots, a_n \in \mathbb{Z}$ . The set of all algebraic integers in a given algebraic number field  $K$  is a ring  $\mathcal{O}(K)$ . It may be shown that, also in  $\mathcal{O}(K)$ , any nonzero proper ideal can be represented as a product of prime ideals and the representation is unique except for the order of the factors. One may also construct the *ideal class group* of  $K$  and show that it is finite, its order being the *class number* of  $K$ .

Some of the motivation for these generalizations came from 'Fermat's last theorem'. Fermat (c. 1640) asserted that the equation  $x^n + y^n = z^n$  has no solutions in positive integers  $x, y, z$  if  $n > 2$ . In Proposition 12 we proved Fermat's assertion for  $n = 3$ . To prove the assertion in general it is sufficient to prove it when  $n = 4$  and when  $n = p$  is an odd prime, since if  $x^{km} + y^{km} = z^{km}$ , then  $(x^k)^m + (y^k)^m = (z^k)^m$ . Fermat himself gave a proof for  $n = 4$ , which is reproduced in Chapter XIII. Proofs for  $n = 3, 5$  and  $7$  were given by Euler (1760–1770), Legendre (1825) and Lamé (1839) respectively.

Kummer (1850) made a remarkable advance beyond this by proving that the assertion holds whenever  $n = p$  is a 'regular' prime. Here a prime  $p$  is said to be *regular* if it does not divide the class number of the *cyclotomic field*  $\mathbb{Q}(\zeta_p)$ , obtained by adjoining to  $\mathbb{Q}$  a  $p$ -th root of unity  $\zeta_p$ . Kummer converted his result into a practical test by further proving that a prime  $p > 3$  is regular if and only if it does not divide the numerator of any of the *Bernoulli numbers*  $B_2, B_4, \dots, B_{p-3}$ .

The only irregular primes less than 100 are 37, 59 and 67. Other methods for dealing with irregular primes were devised by Kummer (1857) and Vandiver (1929). By

1993 Fermat's assertion had been established in this way for all  $n$  less than four million. However, these methods did not lead to a complete proof of 'Fermat's last theorem'. As will be seen in Chapter XIII, a complete solution was first found by Wiles (1995), using quite different methods.

### 3 Multiplicative Functions

We define a function  $f : \mathbb{N} \rightarrow \mathbb{C}$  to be an *arithmetical function*. The set of all arithmetical functions can be given the structure of a commutative ring in the following way.

For any two functions  $f, g : \mathbb{N} \rightarrow \mathbb{C}$ , we define their *convolution* or *Dirichlet product*  $f * g : \mathbb{N} \rightarrow \mathbb{C}$  by

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

Dirichlet multiplication satisfies the usual commutative and associative laws:

**Lemma 24** *For any three functions  $f, g, h : \mathbb{N} \rightarrow \mathbb{C}$ ,*

$$f * g = g * f, \quad f * (g * h) = (f * g) * h.$$

*Proof.* Since  $n/d$  runs through the positive divisors of  $n$  at the same time as  $d$ ,

$$\begin{aligned} f * g(n) &= \sum_{d|n} f(d)g(n/d) \\ &= \sum_{d|n} f(n/d)g(d) = g * f(n). \end{aligned}$$

To prove the associative law, put  $G = g * h$ . Then

$$\begin{aligned} f * G(n) &= \sum_{de=n} f(d)G(e) = \sum_{de=n} f(d) \sum_{d'd''=e} g(d')h(d'') \\ &= \sum_{dd'd''=n} f(d)g(d')h(d''). \end{aligned}$$

Similarly, if we put  $F = f * g$ , we obtain

$$\begin{aligned} F * h(n) &= \sum_{de=n} F(e)h(d) = \sum_{de=n} \sum_{d'd''=e} f(d')g(d'')h(d) \\ &= \sum_{dd'd''=n} f(d')g(d'')h(d). \end{aligned}$$

Hence  $F * h(n) = f * G(n)$ . □

For any two functions  $f, g : \mathbb{N} \rightarrow \mathbb{C}$ , we define their *sum*  $f + g : \mathbb{N} \rightarrow \mathbb{C}$  in the natural way:

$$(f + g)(n) = f(n) + g(n).$$

It is obvious that addition is commutative and associative, and that the distributive law holds:

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

The function  $\delta : \mathbb{N} \rightarrow \mathbb{C}$ , defined by

$$\delta(n) = 1 \text{ or } 0 \quad \text{according as } n = 1 \text{ or } n > 1,$$

acts as an identity element for Dirichlet multiplication:

$$\delta * f = f \quad \text{for every } f : \mathbb{N} \rightarrow \mathbb{C},$$

since

$$\delta * f(n) = \sum_{d|n} \delta(d) f(n/d) = f(n).$$

Thus the set  $\mathcal{A}$  of all arithmetical functions is indeed a commutative ring.

For any function  $f : \mathbb{N} \rightarrow \mathbb{C}$  which is not identically zero, put  $|f| = v(f)^{-1}$ , where  $v(f)$  is the least positive integer  $n$  such that  $f(n) \neq 0$ , and put  $|0| = 0$ . Then

$$|f * g| = |f||g|, |f + g| \leq \max(|f|, |g|) \quad \text{for all } f, g \in \mathcal{A}.$$

Hence the ring  $\mathcal{A}$  of all arithmetical functions is actually an integral domain. It is readily shown that the set of all  $f \in \mathcal{A}$  such that  $|f| < 1$  is an ideal, but not a principal ideal. (Although  $\mathcal{A}$  is not a principal ideal domain, it may be shown that it is a *factorial* domain.)

The next result shows that the functions  $f \in \mathcal{A}$  such that  $|f| = 1$  are the *units* in the ring  $\mathcal{A}$ :

**Lemma 25** *For any function  $f : \mathbb{N} \rightarrow \mathbb{C}$ , there is a function  $f^{-1} : \mathbb{N} \rightarrow \mathbb{C}$  such that  $f^{-1} * f = \delta$  if and only if  $f(1) \neq 0$ . The inverse  $f^{-1}$  is uniquely determined and  $f^{-1}(1)f(1) = 1$ .*

*Proof* Suppose  $g : \mathbb{N} \rightarrow \mathbb{C}$  has the property that  $g * f = \delta$ . Then  $g(1)f(1) = 1$ . Thus  $g(1)$  is non-zero and uniquely determined. If  $n > 1$ , then

$$\sum_{d|n} g(d) f(n/d) = 0.$$

Hence

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

It follows by induction that  $g(n)$  is uniquely determined for every  $n \in \mathbb{N}$ . Conversely, if  $g$  is defined inductively in this way, then  $g * f = \delta$ .  $\square$

It follows from Lemma 25 that the set of all arithmetical functions  $f : \mathbb{N} \rightarrow \mathbb{C}$  such that  $f(1) \neq 0$  is an abelian group under Dirichlet multiplication.

A function  $f : \mathbb{N} \rightarrow \mathbb{C}$  is said to be *multiplicative* if it is not identically zero and if

$$f(mn) = f(m)f(n) \quad \text{for all } m, n \text{ with } (m, n) = 1.$$

It follows that  $f(1) = 1$ , since  $f(n) \neq 0$  for some  $n$  and  $f(n) = f(n)f(1)$ . Any multiplicative function  $f : \mathbb{N} \rightarrow \mathbb{C}$  is uniquely determined by its values at the prime powers, since if

$$m = p_1^{\alpha_1} \cdots p_s^{\alpha_s},$$

where  $p_1, \dots, p_s$  are distinct primes and  $\alpha_1, \dots, \alpha_s \in \mathbb{N}$ , then

$$f(m) = f(p_1^{\alpha_1}) \cdots f(p_s^{\alpha_s}).$$

If

$$m = \prod_p p^{\alpha_p}, \quad n = \prod_p p^{\beta_p},$$

where  $\alpha_p, \beta_p \geq 0$ , then

$$(m, n) = \prod_p p^{\gamma_p}, \quad [m, n] = \prod_p p^{\delta_p},$$

where  $\gamma_p = \min\{\alpha_p, \beta_p\}$  and  $\delta_p = \max\{\alpha_p, \beta_p\}$ . Since either  $\gamma_p = \alpha_p$  and  $\delta_p = \beta_p$ , or  $\gamma_p = \beta_p$  and  $\delta_p = \alpha_p$ , it follows that, for any multiplicative function  $f$  and all  $m, n \in \mathbb{N}$ ,

$$f((m, n))f([m, n]) = \prod_p f(p^{\gamma_p})f(p^{\delta_p}) = \prod_p f(p^{\alpha_p})f(p^{\beta_p}) = f(m)f(n).$$

As we saw in §5 of Chapter II, it follows from Proposition II.4 that Euler's  $\varphi$ -function is multiplicative. Also, the functions  $i : \mathbb{N} \rightarrow \mathbb{C}$  and  $j : \mathbb{N} \rightarrow \mathbb{C}$ , defined by

$$i(n) = 1, \quad j(n) = n \quad \text{for every } n \in \mathbb{N},$$

are obviously multiplicative. Further examples of multiplicative functions can be constructed with the aid of the next two propositions.

**Proposition 26** *If  $f, g : \mathbb{N} \rightarrow \mathbb{C}$  are multiplicative functions, then their Dirichlet product  $h = f * g$  is also multiplicative.*

*Proof* We have

$$h(n) = \sum_{d|n} f(d)g(n/d).$$

Suppose  $n = n'n''$ , where  $n'$  and  $n''$  are relatively prime. Then, by Proposition II.4,

$$\begin{aligned}
h(n) &= \sum_{d'|n', d''|n''} f(d'd'')g(n'n''/d'd'') \\
&= \sum_{d'|n', d''|n''} f(d')f(d'')g(n'/d')g(n''/d'') \\
&= \sum_{d'|n'} f(d')g(n'/d') \sum_{d''|n''} f(d'')g(n''/d'') = h(n')h(n''). \quad \square
\end{aligned}$$

**Proposition 27** *If  $f : \mathbb{N} \rightarrow \mathbb{C}$  is a multiplicative function, then its Dirichlet inverse  $f^{-1} : \mathbb{N} \rightarrow \mathbb{C}$  is also multiplicative.*

*Proof* Assume on the contrary that  $g := f^{-1}$  is not multiplicative and let  $n', n''$  be relatively prime positive integers such that  $g(n'n'') \neq g(n')g(n'')$ . We suppose  $n', n''$  chosen so that the product  $n = n'n''$  is minimal. Since  $f$  is multiplicative,  $f(1) = 1$  and hence  $g(1) = 1$ . Consequently  $n' > 1, n'' > 1$  and

$$0 = \sum_{d'|n'} g(d')f(n'/d') = \sum_{d''|n''} g(d'')f(n''/d'') = \sum_{d|n} g(d)f(n/d).$$

But

$$\begin{aligned}
\sum_{d|n} g(d)f(n/d) &= g(n)f(1) + \sum_{d'|n', d''|n'', d'd'' < n} g(d'd'')f(n'n''/d'd'') \\
&= g(n) + \sum_{d'|n', d''|n'', d'd'' < n} g(d')g(d'')f(n'/d')f(n''/d'') \\
&= g(n) - g(n')g(n'') + \sum_{d'|n'} g(d')f(n'/d') \cdot \sum_{d''|n''} g(d'')f(n''/d'') \\
&= g(n) - g(n')g(n'').
\end{aligned}$$

Thus we have a contradiction.  $\square$

It follows from Propositions 26 and 27 that under Dirichlet multiplication the multiplicative functions form a subgroup of the group of all functions  $f : \mathbb{N} \rightarrow \mathbb{C}$  with  $f(1) \neq 0$ . The further subgroup generated by  $i$  and  $j$  contains some interesting functions. Let  $\tau(n)$  denote the number of positive divisors of  $n$ , and let  $\sigma(n)$  denote the sum of the positive divisors of  $n$ :

$$\tau(n) = \sum_{d|n} 1, \quad \sigma(n) = \sum_{d|n} d.$$

In other words,

$$\tau = i * i, \quad \sigma = i * j,$$

and hence, by Proposition 26,  $\tau$  and  $\sigma$  are multiplicative functions. Thus they are uniquely determined by their values at the prime powers. But if  $p$  is prime and  $\alpha \in \mathbb{N}$ , the divisors of  $p^\alpha$  are  $1, p, \dots, p^\alpha$  and hence

$$\tau(p^\alpha) = \alpha + 1, \quad \sigma(p^\alpha) = (p^{\alpha+1} - 1)/(p - 1).$$

By Proposition II.24, Euler's  $\varphi$ -function satisfies  $i * \varphi = j$ . Thus  $\varphi = i^{-1} * j$ , and Propositions 26 and 27 provide a new proof that Euler's  $\varphi$ -function is multiplicative. Since

$$\tau * \varphi = i * i * \varphi = i * j = \sigma,$$

we also obtain the new relation

$$\sigma(n) = \sum_{d|n} \tau(n/d) \varphi(d).$$

The *Möbius function*  $\mu : \mathbb{N} \rightarrow \mathbb{C}$  is defined to be the Dirichlet inverse  $i^{-1}$ . Thus  $\mu * i = \delta$  or, in other words,

$$\sum_{d|n} \mu(d) = 1 \text{ or } 0 \quad \text{according as } n = 1 \text{ or } n > 1.$$

Instead of this inductive definition, we may explicitly characterize the Möbius function in the following way:

**Proposition 28** *For any  $n \in \mathbb{N}$ ,*

$$\mu(n) = \begin{cases} 1 & \text{if } n = 1, \\ (-1)^s & \text{if } n \text{ is a product of } s \text{ distinct primes,} \\ 0 & \text{if } n \text{ is divisible by the square of a prime.} \end{cases}$$

*Proof.* It is trivial that  $\mu(1) = 1$ . Suppose  $p$  is prime and  $\alpha \in \mathbb{N}$ . Since the divisors of  $p^\alpha$  are  $1, p, \dots, p^\alpha$ , we have

$$\mu(1) + \mu(p) + \dots + \mu(p^\alpha) = 0.$$

Since this holds for all  $\alpha \in \mathbb{N}$ , it follows that  $\mu(p) = -\mu(1) = -1$ , whereas  $\mu(p^\alpha) = 0$  if  $\alpha > 1$ . Since the Möbius function is multiplicative, by Proposition 27, the general formula follows.  $\square$

The function defined by the statement of Proposition 28 had already appeared in work of Euler (1748), but Möbius (1832) discovered the basic property which we have adopted as a definition. From this property we can easily derive the *Möbius inversion formula*:

**Proposition 29** *For any function  $f : \mathbb{N} \rightarrow \mathbb{C}$ , if  $\hat{f} : \mathbb{N} \rightarrow \mathbb{C}$  is defined by*

$$\hat{f}(n) = \sum_{d|n} f(d),$$

*then*

$$f(n) = \sum_{d|n} \hat{f}(d) \mu(n/d) = \sum_{d|n} \hat{f}(n/d) \mu(d).$$

*Furthermore, for any function  $\hat{f} : \mathbb{N} \rightarrow \mathbb{C}$ , there is a unique function  $f : \mathbb{N} \rightarrow \mathbb{C}$  such that  $\hat{f}(n) = \sum_{d|n} f(d)$  for every  $n \in \mathbb{N}$ .*

*Proof* Let  $f : \mathbb{N} \rightarrow \mathbb{C}$  be given and put  $\hat{f} = f * i$ . Then

$$\hat{f} * \mu = f * i * \mu = f * \delta = f.$$

Conversely, let  $\hat{f} : \mathbb{N} \rightarrow \mathbb{C}$  be given and put  $f = \hat{f} * \mu$ . Then  $f * i = \hat{f} * \delta = \hat{f}$ . Moreover, by the first part of the proof, this is the only possible choice for  $f$ .  $\square$

If we apply Proposition 29 to Euler's  $\varphi$ -function then, by Proposition II.24, we obtain the formula

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

In particular, if  $n = p^\alpha$ , where  $p$  is prime and  $\alpha \in \mathbb{N}$ , then

$$\varphi(p^\alpha) = \mu(1)p^\alpha + \mu(p)p^{\alpha-1} = p^\alpha(1 - 1/p).$$

Since  $\varphi$  is multiplicative, we recover in this way the formula

$$\varphi(n) = n \prod_{p|n} (1 - 1/p) \quad \text{for every } n \in \mathbb{N}.$$

The  $\sigma$ -function arises in the study of perfect numbers, to which the Pythagoreans attached much significance. A positive integer  $n$  is said to be *perfect* if it is the sum of its (positive) divisors other than itself, i.e. if  $\sigma(n) = 2n$ .

For example, 6 and 28 are perfect, since

$$6 = 1 + 2 + 3, \quad 28 = 1 + 2 + 4 + 7 + 14.$$

It is an age-old conjecture that there are no odd perfect numbers. However, the even perfect numbers are characterized by the following result:

**Proposition 30** *An even positive integer is perfect if and only if it has the form  $2^t(2^{t+1} - 1)$ , where  $t \in \mathbb{N}$  and  $2^{t+1} - 1$  is prime.*

*Proof* Let  $n$  be any even positive integer and write  $n = 2^t m$ , where  $t \geq 1$  and  $m$  is odd. Then, since  $\sigma$  is multiplicative,  $\sigma(n) = d\sigma(m)$ , where

$$d := \sigma(2^t) = 2^{t+1} - 1.$$

If  $m = d$  and  $d$  is prime, then  $\sigma(m) = 1 + d = 2^{t+1}$  and consequently  $\sigma(n) = 2^{t+1}m = 2n$ .

On the other hand, if  $\sigma(n) = 2n$ , then  $d\sigma(m) = 2^{t+1}m$ . Since  $d$  is odd, it follows that  $m = dq$  for some  $q \in \mathbb{N}$ . Hence

$$\sigma(m) = 2^{t+1}q = (1 + d)q = q + m.$$

Thus  $q$  is the only proper divisor of  $m$ . Hence  $q = 1$  and  $m = d$  is prime.  $\square$

The sufficiency of the condition in Proposition 30 was proved in Euclid's *Elements* (Book IX, Proposition 36). The necessity of the condition was proved over two thousand years later by Euler. The condition is quite restrictive. In the first place, if  $2^m - 1$  is prime for some  $m \in \mathbb{N}$ , then  $m$  must itself be prime. For, if  $m = rs$ , where  $1 < r < m$ , then with  $a = 2^s$  we have

$$2^m - 1 = a^r - 1 = (a - 1)(a^{r-1} + a^{r-2} + \cdots + 1).$$

A prime of the form  $M_p := 2^p - 1$  is said to be a *Mersenne prime* in honour of Mersenne (1644), who gave a list of all primes  $p \leq 257$  for which, he claimed,  $M_p$  was prime. However, he included two values of  $p$  for which  $M_p$  is composite and omitted three values of  $p$  for which  $M_p$  is prime. The correct list is now known to be

$$p = 2, 3, 5, 7, 13, 17, 19, 31, 61, 89, 107, 127.$$

The first four even perfect numbers, namely 6, 28, 496 and 8128, which correspond to the values  $p = 2, 3, 5$  and 7, were known to the ancient Greeks.

That  $M_{11}$  is not prime follows from  $2^{11} - 1 = 2047 = 23 \times 89$ . The factor 23 is not found simply by guesswork. It was already known to Fermat (1640) that if  $p$  is an odd prime, then any divisor of  $M_p$  is congruent to 1 mod  $2p$ . It is sufficient to establish this for prime divisors. But if  $q$  is a prime divisor of  $M_p$ , then  $2^p \equiv 1 \pmod{q}$ . Hence the order of 2 in  $\mathbb{F}_q^\times$  divides  $p$  and, since it is not 1, it must be exactly  $p$ . Hence, by Lemma II.31 with  $G = \mathbb{F}_q^\times$ ,  $p$  divides  $q - 1$ . Thus  $q \equiv 1 \pmod{p}$  and actually  $q \equiv 1 \pmod{2p}$ , since  $q$  is necessarily odd.

The least 39 Mersenne primes are now known. The hunt for more uses thousands of linked personal computers and the following test, which was stated by Lucas (1878), but first completely proved by D.H. Lehmer (1930):

**Proposition 31** Define the sequence  $(S_n)$  recurrently by

$$S_1 = 4, \quad S_{n+1} = S_n^2 - 2 \quad (n \geq 1).$$

Then, for any odd prime  $p$ ,  $M_p := 2^p - 1$  is prime if and only if it divides  $S_{p-1}$ .

*Proof* Put

$$\omega = 2 + \sqrt{3}, \quad \omega' = 2 - \sqrt{3}.$$

Since  $\omega\omega' = 1$ , it is easily shown by induction that

$$S_n = \omega^{2^{n-1}} + \omega'^{2^{n-1}} \quad (n \geq 1).$$

Let  $q$  be a prime and let  $J$  denote the set of all real numbers of the form  $a + b\sqrt{3}$ , where  $a, b \in \mathbb{Z}$ . Evidently  $J$  is a commutative ring. By identifying two elements  $a + b\sqrt{3}$  and  $\tilde{a} + \tilde{b}\sqrt{3}$  of  $J$  when  $a \equiv \tilde{a}$  and  $b \equiv \tilde{b} \pmod{q}$ , we obtain a finite commutative ring  $J_q$  containing  $q^2$  elements. The set  $J_q^\times$  of all invertible elements of  $J_q$  is a commutative group containing at most  $q^2 - 1$  elements, since  $0 \notin J_q^\times$ .

Suppose first that  $M_p$  divides  $S_{p-1}$  and assume that  $M_p$  is composite. If  $q$  is the least prime divisor of  $M_p$ , then  $q^2 \leq M_p$  and  $q \neq 2$ . By hypothesis,

$$\omega^{2^{p-2}} + \omega'^{2^{p-2}} \equiv 0 \pmod{q}.$$

Now consider  $\omega$  and  $\omega'$  as elements of  $J_q$ . By multiplying by  $\omega^{2^{p-2}}$ , we obtain  $\omega^{2^{p-1}} = -1$  and hence  $\omega^{2^p} = 1$ . Thus  $\omega \in J_q^\times$  and the order of  $\omega$  in  $J_q^\times$  is exactly  $2^p$ . Hence

$$2^p \leq q^2 - 1 \leq M_p - 1 = 2^p - 2,$$

which is a contradiction.

Suppose next that  $M_p = q$  is prime. Then  $q \equiv -1 \pmod{8}$ , since  $p \geq 3$ . Since  $(2/q) = (-1)^{(q^2-1)/8}$ , it follows that 2 is a quadratic residue of  $q$ . Thus there exists an integer  $a$  such that

$$a^2 \equiv 2 \pmod{q}.$$

Furthermore  $q \equiv 1 \pmod{3}$ , since  $2^2 \equiv 1$  and hence  $2^{p-1} \equiv 1 \pmod{3}$ . Thus  $q$  is a quadratic residue of 3. Since  $q \equiv -1 \pmod{4}$ , it follows from the law of quadratic reciprocity that 3 is a quadratic nonresidue of  $q$ . Hence, by Euler's criterion (Proposition II.28),

$$3^{(q-1)/2} \equiv -1 \pmod{q}.$$

Consider the element  $\tau = a^{q-2}(1 + \sqrt{3})$  of  $J_q$ . We have

$$\tau^2 = 2^{q-2} \cdot 2\omega = \omega,$$

since  $2^{q-1} \equiv 1 \pmod{q}$ . On the other hand,

$$(1 + \sqrt{3})^q = 1 + 3^{(q-1)/2}\sqrt{3} = 1 - \sqrt{3}$$

and hence

$$\tau^q = a^{q-2}(1 - \sqrt{3}).$$

Consequently,

$$\omega^{(q+1)/2} = \tau^{q+1} = a^{q-2}(1 - \sqrt{3}) \cdot a^{q-2}(1 + \sqrt{3}) = 2^{q-2}(-2) = -1.$$

Multiplying by  $\omega'^{(q+1)/4}$ , we obtain  $\omega^{(q+1)/4} = -\omega'^{(q+1)/4}$ . In other words, since  $(q+1)/4 = 2^{p-2}$ ,

$$S_{p-1} = \omega^{2^{p-2}} + \omega'^{2^{p-2}} \equiv 0 \pmod{q}. \quad \square$$

It is conjectured that there are infinitely many Mersenne primes, and hence infinitely many even perfect numbers. A heuristic argument of Gillies (1964), as modified by Wagstaff (1983), suggests that the number of primes  $p \leq x$  for which  $M_p$  is prime is asymptotic to  $(e^\gamma / \log 2) \log x$ , where  $\gamma$  is Euler's constant (Chapter IX, §4) and thus  $e^\gamma / \log 2 = 2.570 \dots$

We turn now from the primality of  $2^m - 1$  to the primality of  $2^m + 1$ . It is easily seen that if  $2^m + 1$  is prime for some  $m \in \mathbb{N}$ , then  $m$  must be a power of 2. For, if  $m = rs$ , where  $r > 1$  is odd, then with  $a = 2^s$  we have

$$2^m + 1 = a^r + 1 = (a + 1)(a^{r-1} - a^{r-2} + \cdots + 1).$$

Put  $F_n := 2^{2^n} + 1$ . Thus, in particular,

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

Evidently  $F_{n+1} - 2 = (F_n - 2)F_n$ , from which it follows by induction that

$$F_n - 2 = F_0 F_1 \cdots F_{n-1} \quad (n \geq 1).$$

Since  $F_n$  is odd, this implies that  $(F_m, F_n) = 1$  if  $m \neq n$ . As a byproduct, we have a proof that there are infinitely many primes.

It is easily verified that  $F_n$  itself is prime for  $n \leq 4$ . It was conjectured by Fermat that the ‘Fermat numbers’  $F_n$  are all prime. However, this was disproved by Euler, who showed that 641 divides  $F_5$ . In fact

$$641 = 5 \cdot 2^7 + 1 = 5^4 + 2^4.$$

Thus  $5 \cdot 2^7 \equiv -1 \pmod{641}$  and hence  $2^{32} \equiv -5^4 \cdot 2^{28} \equiv -(-1)^4 \equiv -1 \pmod{641}$ .

Fermat may have been as wrong as possible, since no  $F_n$  with  $n > 4$  is known to be prime, although many have been proved to be composite. The Fermat numbers which *are* prime found an unexpected application to the construction of regular polygons by ruler and compass, the only instruments which Euclid allowed himself. It was shown by Gauss, at the age of 19, that a regular polygon of  $m$  sides can be constructed by ruler and compass if the order  $\varphi(m)$  of  $\mathbb{Z}_{(m)}^\times$  is a power of 2. It follows from the formula  $\varphi(p^a) = p^{a-1}(p-1)$ , and the multiplicative nature of Euler’s function, that  $\varphi(m)$  is a power of 2 if and only if  $m$  has the form  $2^k \cdot p_1 \cdots p_s$ , where  $k \geq 0$  and  $p_1, \dots, p_s$  are distinct Fermat primes. (Wantzel (1837) showed that a regular polygon of  $m$  sides cannot be constructed by ruler and compass unless  $m$  has this form.) Gauss’s result, in which he took particular pride, was a forerunner of Galois theory and is today usually established as an application of that theory.

The factor 641 of  $F_5$  is not found simply by guesswork. Indeed, we now show that any divisor of  $F_n$  must be congruent to 1 mod  $2^{n+1}$ . It is sufficient to establish this for prime divisors. But if  $p$  is a prime divisor of  $F_n$ , then  $2^{2^n} \equiv -1 \pmod{p}$  and hence  $2^{2^{n+1}} \equiv 1 \pmod{p}$ . Thus the order of 2 in  $\mathbb{F}_p^\times$  is exactly  $2^{n+1}$ . Hence  $2^{n+1}$  divides  $p-1$  and  $p \equiv 1 \pmod{2^{n+1}}$ .

With a little more effort we can show that any divisor of  $F_n$  must be congruent to 1 mod  $2^{n+2}$  if  $n > 1$ . For, if  $p$  is a prime divisor of  $F_n$  and  $n > 1$ , then  $p \equiv 1 \pmod{8}$  by what we have already proved. Hence, by Proposition II.30, 2 is a quadratic residue of  $p$ . Thus there exists an integer  $a$  such that  $a^2 \equiv 2 \pmod{p}$ . Since  $a^{2^{n+1}} \equiv -1 \pmod{p}$  and  $a^{2^{n+2}} \equiv 1 \pmod{p}$ , the order of  $a$  in  $\mathbb{F}_p^\times$  is exactly  $2^{n+2}$  and hence  $2^{n+2}$  divides  $p-1$ .

It follows from the preceding result that 641 is the first possible candidate for a prime divisor of  $F_5$ , since  $128k + 1$  is not prime for  $k = 1, 3, 4$  and  $257 = F_3$  is relatively prime to  $F_5$ .

The hunt for Fermat primes today uses supercomputers and the following test due to Pépin (1877):

**Proposition 32** *If  $m > 1$ , then  $N := 2^m + 1$  is prime if and only if  $3^{(N-1)/2} + 1$  is divisible by  $N$ .*

*Proof* Suppose first that  $N$  divides  $a^{(N-1)/2} + 1$  for some integer  $a$ . If  $p$  is any prime divisor of  $N$ , then  $a^{(N-1)/2} \equiv -1 \pmod{p}$  and hence  $a^{N-1} \equiv 1 \pmod{p}$ . Thus, since  $p$  is necessarily odd, the order of  $a$  in  $\mathbb{F}_p^\times$  divides  $N - 1 = 2^m$ , but does not divide  $(N - 1)/2 = 2^{m-1}$ . Hence the order of  $a$  must be exactly  $2^m$ . Consequently, by Lemma II.31 with  $G = \mathbb{F}_p^\times$ ,  $2^m$  divides  $p - 1$ . Thus

$$2^m \leq p - 1 \leq N - 1 = 2^m,$$

which implies that  $N = p$  is prime.

To prove the converse we use the law of quadratic reciprocity. Suppose  $N = p$  is prime. Then  $p > 3$ , since  $m > 1$ . From  $2 \equiv -1 \pmod{3}$  we obtain  $p \equiv (-1)^m + 1 \pmod{3}$ . Since  $3 \nmid p$ , it follows that  $p \equiv -1 \pmod{3}$ . Thus  $p$  is a quadratic non-residue of 3. But  $p \equiv 1 \pmod{4}$ , since  $m > 1$ . Consequently, by the law of quadratic reciprocity, 3 is a quadratic non-residue of  $p$ . Hence, by Euler's criterion,  $3^{(p-1)/2} \equiv -1 \pmod{p}$ .  $\square$

By means of Proposition 32 it has been shown that  $F_{14}$  is composite, even though no nontrivial factors are known!

## 4 Linear Diophantine Equations

A *Diophantine equation* is an algebraic equation with integer coefficients of which the integer solutions are required. The name honours Diophantus of Alexandria (3rd century A.D.), who solved many problems of this type, although the surviving books of his *Arithmetica* do not treat the linear problems with which we will be concerned.

We wish to determine integers  $x_1, \dots, x_n$  such that

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= c_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= c_2 \\ &\vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= c_m, \end{aligned}$$

where the coefficients  $a_{jk}$  and the right sides  $c_j$  are given integers ( $1 \leq j \leq m$ ,  $1 \leq k \leq n$ ). We may write the system, in matrix notation, as

$$Ax = c.$$

The problem may also be put geometrically. A nonempty set  $M \subseteq \mathbb{Z}^m$  is said to be a  $\mathbb{Z}$ -*module*, or simply a *module*, if  $a, b \in M$  and  $x, y \in \mathbb{Z}$  imply  $xa + yb \in M$ .

For example, if  $a_1, \dots, a_n$  is a finite subset of  $\mathbb{Z}^m$ , then the set  $M$  of all linear combinations  $x_1a_1 + \cdots + x_na_n$  with  $x_1, \dots, x_n \in \mathbb{Z}$  is a module, the module *generated* by  $a_1, \dots, a_n$ . If we take  $a_1, \dots, a_n$  to be the columns of the matrix  $A$ , then  $M$  is the set of all vectors  $Ax$  with  $x \in \mathbb{Z}^n$  and the system  $Ax = c$  is soluble if and only if  $c \in M$ .

If a module  $M$  is generated by the elements  $a_1, \dots, a_n$ , then it is also generated by the elements  $b_1, \dots, b_n$ , where

$$b_k = u_{1k}a_1 + \dots + u_{nk}a_n \quad (u_{jk} \in \mathbb{Z} : 1 \leq j, k \leq n),$$

if the matrix  $U = (u_{jk})$  is invertible. Here an  $n \times n$  matrix  $U$  of integers is said to be *invertible* if there exists an  $n \times n$  matrix  $U^{-1}$  of integers such that  $U^{-1}U = I_n$  or, equivalently,  $UU^{-1} = I_n$ .

For example, if  $ax + by = 1$ , then the matrix

$$U = \begin{pmatrix} a & b \\ -y & x \end{pmatrix}$$

is invertible, with inverse

$$U^{-1} = \begin{pmatrix} x & -b \\ y & a \end{pmatrix}.$$

It may be shown, although we will not use it, that an  $n \times n$  matrix  $U$  is invertible if and only if its determinant  $\det U$  is a *unit*, i.e.  $\det U = \pm 1$ . Under matrix multiplication, the set of all invertible  $n \times n$  matrices of integers is a group, usually denoted by  $GL_n(\mathbb{Z})$ .

To solve the linear Diophantine system  $Ax = c$  we replace it by a system  $By = c$ , where  $B = AU$  for some invertible matrix  $U$ . The idea is to choose  $U$  so that  $B$  has such a simple form that  $y$  can be determined immediately, and then  $x = Uy$ .

We will use the elementary fact that interchanging two columns of a matrix  $A$ , or adding an integral multiple of one column to another column, is equivalent to postmultiplying  $A$  by a suitable invertible matrix  $U$ . In fact  $U$  is obtained by performing the same column operation on the identity matrix  $I_n$ . In the following discussion ‘matrix’ will mean ‘matrix with entries from  $\mathbb{Z}$ ’.

**Proposition 33** *If  $A = (a_1 \dots a_n)$  is a  $1 \times n$  matrix, then there exists an invertible  $n \times n$  matrix  $U$  such that*

$$AU = (d \ 0 \dots 0)$$

*if and only if  $d$  is a greatest common divisor of  $a_1, \dots, a_n$ .*

*Proof* Suppose first that there exists such a matrix  $U$ . Since

$$A = (d \ 0 \dots 0)U^{-1},$$

$d$  is a common divisor of  $a_1, \dots, a_n$ . On the other hand,

$$d = a_1b_1 + \dots + a_nb_n,$$

where  $b_1, \dots, b_n$  is the first column of  $U$ . Hence any common divisor of  $a_1, \dots, a_n$  divides  $d$ . Thus  $d$  is a greatest common divisor of  $a_1, \dots, a_n$ .

Suppose next that  $a_1, \dots, a_n$  have greatest common divisor  $d$ . Since there is nothing to do if  $n = 1$ , we assume  $n > 1$  and use induction on  $n$ . Then if  $d'$  is the greatest

common divisor of  $a_2, \dots, a_n$ , there exists an invertible  $(n-1) \times (n-1)$  matrix  $V'$  such that

$$(a_2 \cdots a_n)V' = (d' \ 0 \cdots 0).$$

Since  $d$  is the greatest common divisor of  $a_1$  and  $d'$ , there exist integers  $u, v$  such that

$$a_1u + d'v = d.$$

Put  $V = I_1 \oplus V'$  and  $W = W' \oplus I_{n-2}$ , where

$$W' = \begin{pmatrix} u & -d'/d \\ v & a_1/d \end{pmatrix}.$$

Then  $V$  and  $W$  are invertible, and

$$(a_1 \ a_2 \cdots a_n)VW = (a_1 \ d' \ 0 \cdots 0)W = (d \ 0 \cdots 0).$$

Thus we can take  $U = VW$ . □

**Corollary 34** *For any given integers  $a_1, \dots, a_n$ , there exists an invertible  $n \times n$  matrix  $U$  with  $a_1, \dots, a_n$  as its first row if and only if the greatest common divisor of  $a_1, \dots, a_n$  is 1.*

*Proof* An invertible matrix  $U$  has  $a_1, \dots, a_n$  as its first row if and only if

$$(a_1 \ a_2 \cdots a_n) = (1 \ 0 \cdots 0)U. \quad \square$$

If  $U$  is invertible, then its transpose is also invertible. It follows that there exists an invertible  $n \times n$  matrix with  $a_1, \dots, a_n$  as its first column also if and only if the greatest common divisor of  $a_1, \dots, a_n$  is 1.

**Proposition 35** *For any  $m \times n$  matrix  $A$ , there exists an invertible  $n \times n$  matrix  $U$  such that  $B = AU$  has the form*

$$B = (B_1 \ 0),$$

where  $B_1$  is an  $m \times r$  submatrix of rank  $r$ .

*Proof* Let  $A$  have rank  $r$ . If  $r = n$ , there is nothing to do. If  $r < n$  and we denote the columns of  $A$  by  $\mathbf{a}_1, \dots, \mathbf{a}_n$ , then there exist  $x_1, \dots, x_n \in \mathbb{Z}$ , not all zero, such that

$$x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n = \mathbf{0}.$$

Moreover, we may assume that  $x_1, \dots, x_n$  have greatest common divisor 1. Then, by Corollary 34, there exists an invertible  $n \times n$  matrix  $U'$  with  $x_1, \dots, x_n$  as its last column. Hence  $A' := AU'$  has its last column zero. If  $r < n-1$ , we can apply the same argument to the submatrix formed by the first  $n-1$  columns of  $A'$ , and so on until we arrive at a matrix  $B$  of the required form. □

The elements  $\mathbf{b}_1, \dots, \mathbf{b}_r$  of a module  $\mathbf{M}$  are said to be a *basis* for  $\mathbf{M}$  if they generate  $\mathbf{M}$  and are linearly independent, i.e.  $x_1\mathbf{b}_1 + \dots + x_r\mathbf{b}_r = \mathbf{O}$  for some  $x_1, \dots, x_r \in \mathbb{Z}$  implies that  $x_1 = \dots = x_r = 0$ . If  $\mathbf{O}$  is the only element of  $\mathbf{M}$ , we say also that  $\mathbf{O}$  is a basis for  $\mathbf{M}$ .

In geometric terms, Proposition 35 says that any finitely generated module  $\mathbf{M} \subseteq \mathbb{Z}^m$  has a finite basis, and that a finite set of generators is a basis if and only if its elements are linearly independent over  $\mathbb{Q}$ . Hence any two bases have the same cardinality.

**Proposition 36** *For any  $m \times n$  matrix  $A$ , the set  $N$  of all  $\mathbf{x} \in \mathbb{Z}^n$  such that  $A\mathbf{x} = \mathbf{O}$  is a module with a finite basis.*

*Proof* It is evident that  $N$  is a module. By Proposition 35, there exists an invertible  $n \times n$  matrix  $U$  such that  $AU = B = (B_1 \mathbf{O})$ , where  $B_1$  is an  $m \times r$  submatrix of rank  $r$ . Hence  $B\mathbf{y} = \mathbf{O}$  if and only if the first  $r$  coordinates of  $\mathbf{y}$  vanish. By taking  $\mathbf{y}$  to be the vector with  $k$ -th coordinate 1 and all other coordinates 0, for each  $k$  such that  $r < k \leq n$ , we see that the equation  $B\mathbf{y} = \mathbf{O}$  has  $n - r$  linearly independent solutions  $\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n-r)}$  such that all solutions are given by

$$\mathbf{y} = z_1\mathbf{y}^{(1)} + \dots + z_{n-r}\mathbf{y}^{(n-r)},$$

where  $z_1, \dots, z_{n-r}$  are arbitrary integers. If we put  $\mathbf{x}^{(j)} = U\mathbf{y}^{(j)}$ , it follows that  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n-r)}$  are a basis for the module  $N$ .  $\square$

An  $m \times n$  matrix  $B = (b_{jk})$  will be said to be in *echelon form* if the following two conditions are satisfied:

- (i)  $b_{jk} = 0$  for all  $j$  if  $k > r$ ;
- (ii)  $b_{jk} \neq 0$  for some  $j$  if  $k \leq r$  and, if  $m_k$  is the least such  $j$ , then  $1 \leq m_1 < m_2 < \dots < m_r \leq m$ .

Evidently  $r = \text{rank } B$ .

**Proposition 37** *For any  $m \times n$  matrix  $A$ , there exists an invertible  $n \times n$  matrix  $U$  such that  $B = AU$  is in echelon form.*

*Proof* By Proposition 35, we may suppose that  $A$  has the form  $(A_1 \mathbf{O})$ , where  $A_1$  is an  $m \times r$  submatrix of rank  $r$ , and by replacing  $A_1$  by  $A$  we may suppose that  $A$  itself has rank  $n$ . We are going to show that there exists an invertible  $n \times n$  matrix  $U$  such that, if  $AU = B = (b_{jk})$ , then  $b_{jk} = 0$  for all  $j < k$ .

If  $m = 1$ , this follows from Proposition 33. We assume  $m > 1$  and use induction on  $m$ . Then the first  $m - 1$  rows of  $A$  may be assumed to have already the required triangular form. If  $n \leq m$ , there is nothing more to do. If  $n > m$ , we can take  $U = I_{m-1} \oplus U'$ , where  $U'$  is an invertible  $(n - m + 1) \times (n - m + 1)$  matrix such that

$$(a_{m,m} \ a_{m,m+1} \ \dots \ a_{m,n})U' = (a' \ 0 \ \dots \ 0).$$

Replacing  $B$  by  $A$ , we now suppose that for  $A$  itself we have  $a_{jk} = 0$  for all  $j < k$ . Since  $A$  still has rank  $n$ , each column of  $A$  contains a nonzero entry. If the first nonzero entry in the  $k$ -th column appears in the  $m_k$ -th row, then  $m_k \geq k$ . By permuting the columns, if necessary, we may suppose in addition that  $m_1 \leq m_2 \leq \dots \leq m_n$ .

Suppose  $m_1 = m_2$ . Let  $a$  and  $b$  be the entries in the  $m_1$ -th row of the first and second columns, and let  $d$  be their greatest common divisor. Then  $d \neq 0$  and there exist integers  $u, v$  such that  $au + bv = d$ . If we put  $U = V \oplus I_{n-2}$ , where

$$V = \begin{pmatrix} u & -b/d \\ v & a/d \end{pmatrix},$$

then  $U$  is invertible. Moreover, the last  $n - 2$  columns of  $B = AU$  are the same as in  $A$  and the first  $m_1 - 1$  entries of the first two columns are still zero. However,  $b_{m_1 1} = d$  and  $b_{m_1 2} = 0$ . By permuting the last  $n - 1$  columns, if necessary, we obtain a matrix  $A'$ , of the same form as  $A$ , with  $m'_1 \leq m'_2 \leq \dots \leq m'_n$ , where  $m'_1 = m_1$  and  $m'_2 + \dots + m'_n > m_2 + \dots + m_n$ .

By repeating this process finitely many times, we will obtain a matrix in echelon form.  $\square$

**Corollary 38** *If  $A$  is an  $m \times n$  matrix of rank  $m$ , then there exists an invertible  $n \times n$  matrix  $U$  such that  $AU = B = (b_{jk})$ , where*

$$b_{jj} \neq 0, \quad b_{jk} = 0 \text{ if } j < k \quad (1 \leq j \leq m, \quad 1 \leq k \leq n).$$

Before proceeding further we consider the uniqueness of the echelon form. Let  $T = (t_{jk})$  be any  $r \times r$  matrix which is lower triangular and invertible, i.e.  $t_{jk} = 0$  if  $j < k$  and the diagonal elements  $t_{jj}$  are units. It is easily seen that if  $U = T \oplus I_{n-r}$ , and if  $B$  is an echelon form for a matrix  $A$  with rank  $r$ , then  $BU$  is also an echelon form for  $A$ . We will show that all possible echelon forms for  $A$  are obtained in this way.

Suppose  $B' = BU$  is in echelon form, for some invertible  $n \times n$  matrix  $U$ , and write

$$B = (B_1 \ O),$$

where  $B_1$  is an  $m \times r$  submatrix. If

$$U = \begin{pmatrix} U_1 & U_2 \\ U_3 & U_4 \end{pmatrix},$$

then from  $(B_1 \ O)U = (B'_1 \ O)$  we obtain  $U_2 = O$ , since  $B_1 U_2 = O$  and  $B_1$  has rank  $r$ . Consequently  $U_1$  is invertible and we can equally well take  $U_3 = O$ ,  $U_4 = I$ . Let  $\mathbf{b}_1, \dots, \mathbf{b}_r$  be the columns of  $B_1$  and  $\mathbf{b}'_1, \dots, \mathbf{b}'_r$  the columns of  $B'_1$ . If  $U_1 = (t_{jk})$ , then

$$\mathbf{b}'_k = t_{1k}\mathbf{b}_1 + \dots + t_{rk}\mathbf{b}_r \quad (1 \leq k \leq r).$$

Taking  $k = 1$ , we obtain  $m'_1 \geq m_1$  and so, by symmetry,  $m'_1 = m_1$ . Since  $m'_k > m'_1$  for  $k > 1$ , it follows that  $t_{1k} = 0$  for  $k > 1$ . Taking  $k = 2$ , we now obtain in the same way  $m'_2 = m_2$ . Proceeding in this manner, we see that  $U_1$  is a lower triangular matrix.

We return now to the linear Diophantine equation

$$A\mathbf{x} = \mathbf{c}.$$

The set of all  $\mathbf{c} \in \mathbb{Z}^m$  for which there exists a solution  $\mathbf{x} \in \mathbb{Z}^n$  is evidently a module  $\mathbf{L} \subseteq \mathbb{Z}^m$ . If  $U$  is an invertible matrix such that  $B = AU$  is in echelon form, then  $\mathbf{x}$  is a solution of the given system if and only if  $\mathbf{y} = U^{-1}\mathbf{x}$  is a solution of the transformed system

$$By = c.$$

But the latter system is soluble if and only if  $c$  is an integral linear combination of the first  $r$  columns  $b_1, \dots, b_r$  of  $B$ . Since  $b_1, \dots, b_r$  are linearly independent, they form a basis for  $L$ .

To determine if a given system  $Ax = c$  is soluble, we may use the customary methods of linear algebra over the field  $\mathbb{Q}$  of rational numbers to test if  $c$  is linearly dependent on  $b_1, \dots, b_r$ ; then express it as a linear combination of  $b_1, \dots, b_r$ , and finally check that the coefficients  $y_1, \dots, y_r$  are all integers. The solutions of the original system are given by  $x = Uy$ , where  $y$  is any vector in  $\mathbb{Z}^n$  whose first  $r$  coordinates are  $y_1, \dots, y_r$ .

If  $M_1$  and  $M_2$  are modules in  $\mathbb{Z}^m$ , their *intersection*  $M_1 \cap M_2$  is again a module. The set of all  $a \in \mathbb{Z}^m$  such that  $a = a_1 + a_2$  for some  $a_1 \in M_1$  and  $a_2 \in M_2$  is also a module, which will be denoted by  $M_1 + M_2$  and called the *sum* of  $M_1$  and  $M_2$ . If  $M_1$  and  $M_2$  are finitely generated, then  $M_1 + M_2$  is evidently finitely generated. We will show that  $M_1 \cap M_2$  is also finitely generated.

Since  $M_1 + M_2$  is a finitely generated module in  $\mathbb{Z}^m$ , it has a basis  $a_1, \dots, a_n$ . Since  $M_1$  and  $M_2$  are contained in  $M_1 + M_2$ , their generators  $b_1, \dots, b_p$  and  $c_1, \dots, c_q$  have the form

$$b_i = \sum_{k=1}^n u_{ki} a_k,$$

$$c_j = \sum_{k=1}^n v_{kj} a_k,$$

for some  $u_{ki}, v_{kj} \in \mathbb{Z}$ . Then  $a \in M_1 \cap M_2$  if and only if

$$a = \sum_{i=1}^p y_i b_i = \sum_{j=1}^q z_j c_j$$

for some  $y_i, z_j \in \mathbb{Z}$ . Since  $a_1, \dots, a_n$  is a basis for  $M_1 + M_2$ , this is equivalent to

$$\sum_{i=1}^p u_{ki} y_i = \sum_{j=1}^q v_{kj} z_j$$

or, in matrix notation,  $By = Cz$ . But this is equivalent to the homogeneous system  $Ax = O$ , where

$$A = (B - C), \quad x = \begin{pmatrix} y \\ z \end{pmatrix},$$

and by Proposition 36 the module of solutions of this system has a finite basis.

Suppose the modules  $M_1, M_2 \subseteq \mathbb{Z}^m$  are generated by the columns of the  $m \times n_1$ ,  $m \times n_2$  matrices  $A_1, A_2$ . Evidently  $M_2$  is a submodule of  $M_1$  if and only if each column of  $A_2$  can be expressed as a linear combination of the columns of  $A_1$ , i.e. if and only if there exists an  $n_1 \times n_2$  matrix  $X$  such that

$$A_1 X = A_2.$$

We say in this case that  $A_1$  is a *left divisor* of  $A_2$ , or that  $A_2$  is a *right multiple* of  $A_1$ .

We may also define greatest common divisors and least common multiples for matrices. An  $m \times p$  matrix  $D$  is a *greatest common left divisor* of  $A_1$  and  $A_2$  if it is a left divisor of both  $A_1$  and  $A_2$ , and if every left divisor  $C$  of both  $A_1$  and  $A_2$  is also a left divisor of  $D$ . An  $m \times q$  matrix  $H$  is a *least common right multiple* of  $A_1$  and  $A_2$  if it is a right multiple of both  $A_1$  and  $A_2$ , and if every right multiple  $G$  of both  $A_1$  and  $A_2$  is also a right multiple of  $H$ . It will now be shown that these objects exist and have simple geometrical interpretations.

Let  $\mathbf{M}_1, \mathbf{M}_2$  be the modules defined by the matrices  $A_1, A_2$ . We will show that if the sum  $\mathbf{M}_1 + \mathbf{M}_2$  is defined by the matrix  $D$ , then  $D$  is a greatest common left divisor of  $A_1$  and  $A_2$ . In fact  $D$  is a common left divisor of  $A_1$  and  $A_2$ , since  $\mathbf{M}_1$  and  $\mathbf{M}_2$  are contained in  $\mathbf{M}_1 + \mathbf{M}_2$ . On the other hand, any common left divisor  $C$  of  $A_1$  and  $A_2$  defines a module which contains  $\mathbf{M}_1 + \mathbf{M}_2$ , since it contains both  $\mathbf{M}_1$  and  $\mathbf{M}_2$ , and so  $C$  is a left divisor of  $D$ .

A similar argument shows that if the intersection  $\mathbf{M}_1 \cap \mathbf{M}_2$  is defined by the matrix  $H$ , then  $H$  is a least common right multiple of  $A_1$  and  $A_2$ .

The sum  $\mathbf{M}_1 + \mathbf{M}_2$  is defined, in particular, by the block matrix  $(A_1 \ A_2)$ . There exists an invertible  $(n_1 + n_2) \times (n_1 + n_2)$  matrix  $U$  such that

$$(A_1 \ A_2)U = (D' \ O),$$

where  $D'$  is an  $m \times r$  submatrix of rank  $r$ . If

$$U = \begin{pmatrix} U_1 & U_2 \\ U_3 & U_4 \end{pmatrix},$$

is the corresponding partition of  $U$ , then

$$A_1 U_1 + A_2 U_3 = D'.$$

On the other hand,

$$(A_1 \ A_2) = (D' \ O)U^{-1}.$$

If

$$U^{-1} = \begin{pmatrix} V_1 & V_2 \\ V_3 & V_4 \end{pmatrix}$$

is the corresponding partition of  $U^{-1}$ , then

$$A_1 = D'V_1, \quad A_2 = D'V_2.$$

Thus  $D'$  is a common left divisor of  $A_1$  and  $A_2$ , and the previous relation implies that it is a greatest common left divisor. It follows that *any* greatest common left divisor  $D$  of  $A_1$  and  $A_2$  has a right 'Bézout' representation  $D = A_1 X_1 + A_2 X_2$ .

We may also define coprimeness for matrices. Two matrices  $A_1, A_2$  of size  $m \times n_1, m \times n_2$  are *left coprime* if  $I_m$  is a greatest common left divisor. If  $\mathbf{M}_1, \mathbf{M}_2$  are the modules defined by  $A_1, A_2$ , this means that  $\mathbf{M}_1 + \mathbf{M}_2 = \mathbb{Z}^m$ . The definition may also be reformulated in several other ways.

**Proposition 39** For any  $m \times n$  matrix  $A$ , the following conditions are equivalent:

- (i) for some, and hence every, partition  $A = (A_1 \ A_2)$ , the submatrices  $A_1$  and  $A_2$  are left coprime;
- (ii) there exists an  $n \times m$  matrix  $A^\dagger$  such that  $AA^\dagger = I_m$ ;
- (iii) there exists an  $(n - m) \times n$  matrix  $A^c$  such that

$$\begin{pmatrix} A \\ A^c \end{pmatrix}$$

is invertible;

- (iv) there exists an invertible  $n \times n$  matrix  $V$  such that  $AV = (I_m \ 0)$ .

*Proof* If  $A = (A_1 \ A_2)$  for some left coprime matrices  $A_1, A_2$ , then there exist  $X_1, X_2$  such that  $A_1X_1 + A_2X_2 = I_m$  and hence (ii) holds. On the other hand, if (ii) holds then, for any partition  $A = (A_1 \ A_2)$ , there exist  $X_1, X_2$  such that  $A_1X_1 + A_2X_2 = I_m$  and hence  $A_1, A_2$  are left coprime.

Thus (i)  $\Leftrightarrow$  (ii). Suppose now that (ii) holds. Then  $A$  has rank  $m$  and hence there exists an invertible  $n \times n$  matrix  $U$  such that  $A = (D \ 0)U$ , where the  $m \times m$  matrix  $D$  is non-singular. In fact  $D$  is invertible, since  $AA^\dagger = I_m$  implies that  $D$  is a left divisor of  $I_m$ . Consequently, by changing  $U$ , we may assume  $D = I_m$ . If we now take  $A^c = (0 \ I_{n-m})U$ , we see that (ii)  $\Rightarrow$  (iii).

It is obvious that (iii)  $\Rightarrow$  (iv) and that (iv)  $\Rightarrow$  (ii).  $\square$

We now consider the extension of these results to other rings besides  $\mathbb{Z}$ . Let  $R$  be an arbitrary ring. A nonempty set  $M \subseteq R^m$  is said to be an  $R$ -module if  $a, b \in M$  and  $x, y \in R$  imply  $xa + yb \in M$ . The module  $M$  is *finitely generated* if it contains elements  $a_1, \dots, a_n$  such that every element of  $M$  has the form  $x_1a_1 + \dots + x_na_n$  for some  $x_1, \dots, x_n \in R$ .

It is easily seen that if  $R$  is a *Bézout domain*, then the whole of the preceding discussion in this section remains valid if ‘module’ is interpreted to mean ‘ $R$ -module’ and ‘matrix’ to mean ‘matrix with entries from  $R$ ’. In particular, we may take  $R = K[t]$  to be the ring of all polynomials in one indeterminate with coefficients from an arbitrary field  $K$ . However, both  $\mathbb{Z}$  and  $K[t]$  are principal ideal domains. In this case further results may be obtained.

**Proposition 40** If  $R$  is a principal ideal domain and  $M$  a finitely generated  $R$ -module, then any submodule  $L$  of  $M$  is also finitely generated. Moreover, if  $M$  is generated by  $n$  elements, so also is  $L$ .

*Proof* Suppose  $M$  is generated by  $a_1, \dots, a_n$ . If  $n = 1$ , then any  $b \in L$  has the form  $b = xa_1$  for some  $x \in R$  and the set of all  $x$  which appear in this way is an ideal of  $R$ . Since  $R$  is a principal ideal domain, it follows that  $L$  is generated by a single element  $b_1$ , where  $b_1 = x'a_1$  for some  $x' \in R$ .

Suppose now that  $n > 1$  and that, for each  $m < n$ , any submodule of a module generated by  $m$  elements is also generated by  $m$  elements. Any  $b \in L$  has the form

$$b = x_1a_1 + \dots + x_na_n$$

for some  $x_1, \dots, x_n \in R$  and the set of all  $x_1$  which appear in this way is an ideal of  $R$ . Since  $R$  is a principal ideal domain, it follows that there is a fixed  $b_1 \in L$  such

that  $\mathbf{b} = y_1 \mathbf{b}_1 + \mathbf{b}'$  for some  $y_1 \in R$  and some  $\mathbf{b}'$  in the module  $\mathbf{M}'$  generated by  $\mathbf{a}_2, \dots, \mathbf{a}_n$ . The set of all  $\mathbf{b}'$  which appear in this way is a submodule  $\mathbf{L}'$  of  $\mathbf{M}'$ . By the induction hypothesis,  $\mathbf{L}'$  is generated by  $n - 1$  elements and hence  $\mathbf{L}$  is generated by  $n$  elements.  $\square$

Just as it is useful to define vector spaces abstractly over an arbitrary field  $K$ , so it is useful to define modules abstractly over an arbitrary ring  $R$ . An abelian group  $\mathbf{M}$ , with the group operation denoted by  $+$ , is said to be an  $R$ -module if, with any  $\mathbf{a} \in \mathbf{M}$  and any  $x \in R$ , there is associated an element  $x\mathbf{a} \in \mathbf{M}$  so that the following properties hold, for all  $\mathbf{a}, \mathbf{b} \in \mathbf{M}$  and all  $x, y \in R$ :

- (i)  $x(\mathbf{a} + \mathbf{b}) = x\mathbf{a} + x\mathbf{b}$ ,
- (ii)  $(x + y)\mathbf{a} = x\mathbf{a} + y\mathbf{a}$ ,
- (iii)  $(xy)\mathbf{a} = x(y\mathbf{a})$ ,
- (iv)  $1\mathbf{a} = \mathbf{a}$ .

The proof of Proposition 40 remains valid for modules in this abstract sense. However, a finitely generated module need not now have a basis. For, even if it is generated by a single element  $\mathbf{a}$ , we may have  $x\mathbf{a} = \mathbf{0}$  for some nonzero  $x \in R$ . Nevertheless, we are going to show that, if  $R$  is a principal ideal domain, all finitely generated  $R$ -modules can be completely characterized.

Let  $R$  be a principal ideal domain and  $\mathbf{M}$  a finitely generated  $R$ -module, with generators  $\mathbf{a}_1, \dots, \mathbf{a}_n$ , say. The set  $N$  of all  $\mathbf{x} = (x_1, \dots, x_n) \in R^n$  such that

$$x_1 \mathbf{a}_1 + \dots + x_n \mathbf{a}_n = \mathbf{0}$$

is evidently a module in  $R^n$ . Hence  $N$  is finitely generated, by Proposition 40. The given module  $\mathbf{M}$  is isomorphic to the quotient module  $R^n/N$ .

Let  $\mathbf{f}_1, \dots, \mathbf{f}_m$  be a set of generators for  $N$  and let  $\mathbf{e}_1, \dots, \mathbf{e}_n$  be a basis for  $R^n$ . Then

$$\mathbf{f}_j = a_{j1}\mathbf{e}_1 + \dots + a_{jn}\mathbf{e}_n \quad (1 \leq j \leq m),$$

for some  $a_{jk} \in R$ . The module  $\mathbf{M}$  is completely determined by the matrix  $A = (a_{jk})$ . However, we can change generators and change bases.

If we put

$$\mathbf{f}'_i = v_{i1}\mathbf{f}_1 + \dots + v_{im}\mathbf{f}_m \quad (1 \leq i \leq m),$$

where  $V = (v_{ij})$  is an invertible  $m \times m$  matrix, then  $\mathbf{f}'_1, \dots, \mathbf{f}'_m$  is also a set of generators for  $N$ . If we put

$$\mathbf{e}_k = u_{k1}\mathbf{e}'_1 + \dots + u_{kn}\mathbf{e}'_n \quad (1 \leq k \leq n),$$

where  $U = (u_{k\ell})$  is an invertible  $n \times n$  matrix, then  $\mathbf{e}'_1, \dots, \mathbf{e}'_n$  is also a basis for  $R^n$ . Moreover

$$\mathbf{f}'_i = b_{i1}\mathbf{e}'_1 + \dots + b_{in}\mathbf{e}'_n \quad (1 \leq i \leq m),$$

where the  $m \times n$  matrix  $B = (b_{i\ell})$  is given by  $B = VAU$ .

The idea is to choose  $V$  and  $U$  so that  $B$  is as simple as possible. This is made precise in the next proposition, first proved by H.J.S. Smith (1861) for  $R = \mathbb{Z}$ . The corresponding matrix  $S$  is known as the *Smith normal form* of  $A$ .

**Proposition 41** *Let  $R$  be a principal ideal domain and let  $A$  be an  $m \times n$  matrix with entries from  $R$ . If  $A$  has rank  $r$ , then there exist invertible  $m \times m$ ,  $n \times n$  matrices  $V$ ,  $U$  with entries from  $R$  such that  $S = VAU$  has the form*

$$S = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix},$$

where  $D = \text{diag}[d_1, \dots, d_r]$  is a diagonal matrix with nonzero entries  $d_i$  and  $d_i | d_j$  for  $1 \leq i \leq j \leq r$ .

*Proof* We show first that it is enough to obtain a matrix which satisfies all the requirements except the divisibility conditions for the  $d$ 's.

If  $a, b$  are nonzero elements of  $R$  with greatest common divisor  $d$ , then there exist  $u, v \in R$  such that  $au + bv = d$ . It is easily verified that

$$\begin{pmatrix} 1 & 1 \\ -bv/d & au/d \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \begin{pmatrix} u & -b/d \\ v & a/d \end{pmatrix} = \begin{pmatrix} d & 0 \\ 0 & ab/d \end{pmatrix},$$

and the outside matrices on the left-hand side are both invertible. By applying this process finitely many times, a non-singular diagonal matrix  $D' = \text{diag}[d'_1, \dots, d'_r]$  may be transformed into a non-singular diagonal matrix  $D = \text{diag}[d_1, \dots, d_r]$  which satisfies  $d_i | d_j$  for  $1 \leq i \leq j \leq r$ .

Consider now an arbitrary matrix  $A$ . By applying Proposition 35 to the transpose of  $A$ , we may reduce the problem to the case where  $A$  has rank  $m$  and then, by Corollary 38, we may suppose further that  $a_{jj} \neq 0, a_{jk} = 0$  for all  $j < k$ .

It is now sufficient to show that, for any  $2 \times 2$  matrix

$$A = \begin{pmatrix} a & 0 \\ b & c \end{pmatrix},$$

with nonzero entries  $a, b, c$ , there exist invertible  $2 \times 2$  matrices  $U, V$  such that  $VAU$  is a diagonal matrix. Moreover, we need only prove this when the greatest common divisor  $(a, b, c) = 1$ . But then there exists  $q \in R$  such that  $(a, b+qc) = 1$ . In fact, take  $q$  to be the product of the distinct primes which divide  $a$  but not  $b$ . For any prime divisor  $p$  of  $a$ , if  $p|b$ , then  $p \nmid c$ ,  $p \nmid q$  and hence  $p \nmid (b+qc)$ ; if  $p \nmid b$ , then  $p|q$  and again  $p \nmid (b+qc)$ .

If we put  $b' = b + qc$ , then there exist  $x, y \in R$  such that  $ax + b'y = 1$ , and hence  $ax + by = 1 - qcy$ . It is easily verified that

$$\begin{pmatrix} x & y \\ -b' & a \end{pmatrix} \begin{pmatrix} a & 0 \\ b & c \end{pmatrix} \begin{pmatrix} 1 & -cy \\ q & 1 - qcy \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & ac \end{pmatrix},$$

and the outside matrices on the left-hand side are both invertible. □

In the important special case  $R = \mathbb{Z}$ , there is a more constructive proof of Proposition 41. Obviously we may suppose  $A \neq O$ . By interchanges of rows and columns we can arrange that  $a_{11}$  is the nonzero entry of  $A$  with minimum absolute value. If there is an entry  $a_{1k}$  ( $k > 1$ ) in the first row which is not divisible by  $a_{11}$ , then we can write  $a_{1k} = za_{11} + a'_{1k}$ , where  $z, a'_{1k} \in \mathbb{Z}$  and  $|a'_{1k}| < |a_{11}|$ . By subtracting  $z$  times the first column from the  $k$ -th column we replace  $a_{1k}$  by  $a'_{1k}$ . Thus we obtain a new matrix  $A$  in which the minimum absolute value of the nonzero entries has been reduced.

On the other hand, if  $a_{11}$  divides  $a_{1k}$  for all  $k > 1$  then, by subtracting multiples of the first column from the remaining columns, we can arrange that  $a_{1k} = 0$  for all  $k > 1$ . If there is now an entry  $a_{j1}$  ( $j > 1$ ) in the first column which is not divisible by  $a_{11}$  then, by subtracting a multiple of the first row from the  $j$ -th row, the minimum absolute value of the nonzero entries can again be reduced. Otherwise, by subtracting multiples of the first row from the remaining rows, we can bring  $A$  to the form

$$\begin{pmatrix} a_{11} & O \\ O & A' \end{pmatrix}.$$

Since  $A \neq O$  and the minimum absolute value of the nonzero entries cannot be reduced indefinitely, we must in any event arrive at a matrix of this form after a finite number of steps. The same procedure can now be applied to the submatrix  $A'$ , and so on until we obtain a matrix

$$\begin{pmatrix} D' & O \\ O & O \end{pmatrix},$$

where  $D'$  is a diagonal matrix with the same rank as  $A$ . As in the first part of the proof of Proposition 41, we can now replace  $D'$  by a diagonal matrix  $D$  which satisfies the divisibility conditions.

Clearly this constructive proof is also valid for any Euclidean domain  $R$  and, in particular, for the polynomial ring  $R = K[t]$ , where  $K$  is an arbitrary field.

It will now be shown that the Smith normal form of a matrix  $A$  is uniquely determined, apart from replacing each  $d_k$  by an arbitrary unit multiple. For, if  $S'$  is another Smith normal form, then  $S' = V'SU'$  for some invertible  $m \times m$ ,  $n \times n$  matrices  $V'$ ,  $U'$ . Since  $d_1$  divides all entries of  $S$ , it also divides all entries of  $S'$ . In particular,  $d_1 | d'_1$ . In the same way  $d'_1 | d_1$  and hence  $d'_1$  is a unit multiple of  $d_1$ . To show that  $d'_k$  is a unit multiple of  $d_k$ , also for  $k > 1$ , it is quickest to use determinants (Chapter V, §1). Since  $d_1 \cdots d_k$  divides all  $k \times k$  subdeterminants or *minors* of  $S$ , it also divides all  $k \times k$  minors of  $S'$ . In particular,  $d_1 \cdots d_k | d'_1 \cdots d'_k$ . Similarly,  $d'_1 \cdots d'_k | d_1 \cdots d_k$  and hence  $d'_1 \cdots d'_k$  is a unit multiple of  $d_1 \cdots d_k$ . The conclusion now follows by induction on  $k$ .

The products  $\Delta_k := d_1 \cdots d_k$  ( $1 \leq k \leq r$ ) are known as the *invariant factors* of the matrix  $A$ . A similar argument to that in the preceding paragraph shows that  $\Delta_k$  is the greatest common divisor of all  $k \times k$  minors of  $A$ .

Two  $m \times n$  matrices  $A, B$  are said to be *equivalent* if there exist invertible  $m \times m$ ,  $n \times n$  matrices  $V, U$  such that  $B = VAU$ . Since equivalence is indeed an 'equivalence relation', the uniqueness of the Smith normal form implies that two  $m \times n$  matrices  $A, B$  are equivalent if and only if they have the same rank and the same invariant factors.

We return now from matrices to modules. Let  $R$  be a principal ideal domain and  $M$  a finitely generated  $R$ -module, with generators  $a_1, \dots, a_n$ . The Smith normal form tells us that  $M$  has generators  $a'_1, \dots, a'_n$ , where

$$a_k = u_{k1}a'_1 + \cdots + u_{kn}a'_n \quad (1 \leq k \leq n)$$

for some invertible matrix  $U = (u_{k\ell})$ , such that  $d_k a'_k = O$  ( $1 \leq k \leq r$ ). Moreover,

$$x_1 a'_1 + \cdots + x_n a'_n = O$$

implies  $x_k = y_k d_k$  for some  $y_k \in R$  if  $1 \leq k \leq r$  and  $x_k = 0$  if  $r < k \leq n$ . In particular,  $x_k a'_k = 0$  for  $1 \leq k \leq n$ , and thus the module  $M$  is the direct sum of the submodules  $M'_1, \dots, M'_n$  generated by  $a'_1, \dots, a'_n$  respectively.

If  $N_k$  denotes the set of all  $x \in R$  such that  $xa'_k = 0$ , then  $N_k$  is the principal ideal of  $R$  generated by  $d_k$  for  $1 \leq k \leq r$  and  $N_k = \{0\}$  for  $r < k \leq n$ . The divisibility conditions on the  $d$ 's imply that  $N_{k+1} \subseteq N_k$  ( $1 \leq k < r$ ). If  $N_k = R$  for some  $k$ , then  $a'_k$  contributes nothing as a generator and may be omitted.

Evidently the submodule  $M'$  generated by  $a'_1, \dots, a'_r$  consists of all  $a \in M$  such that  $xa = 0$  for some nonzero  $x \in R$ , and the submodule  $M''$  generated by  $a'_{r+1}, \dots, a'_n$  has  $a'_{r+1}, \dots, a'_n$  as a basis. Thus we have proved the *structure theorem for finitely generated modules over a principal ideal domain*:

**Proposition 42** *Let  $R$  be a principal ideal domain and  $M$  a finitely generated  $R$ -module. Then  $M$  is the direct sum of two submodules  $M'$  and  $M''$ , where  $M'$  consists of all  $a \in M$  such that  $xa = 0$  for some nonzero  $x \in R$  and  $M''$  has a finite basis.*

*Moreover,  $M'$  is the direct sum of  $s$  submodules  $Ra_1, \dots, Ra_s$ , such that*

$$0 \subset N_s \subseteq \dots \subseteq N_1 \subset R,$$

*where  $N_k$  is the ideal consisting of all  $x \in R$  such that  $xa_k = 0$  ( $1 \leq k \leq s$ ).*

The uniquely determined submodule  $M'$  is called the *torsion submodule* of  $M$ . The *free submodule*  $M''$  is not uniquely determined, although the number of elements in a basis is uniquely determined. Of course, for a particular  $M$  one may have  $M' = \{0\}$  or  $M'' = \{0\}$ .

Any abelian group  $A$ , with the group operation denoted by  $+$ , may be regarded as a  $\mathbb{Z}$ -module by defining  $na$  to be the sum  $a + \dots + a$  with  $n$  summands if  $n \in \mathbb{N}$ , to be  $0$  if  $n = 0$ , and to be  $-(a + \dots + a)$  with  $-n$  summands if  $-n \in \mathbb{N}$ . The structure theorem in this case becomes the *structure theorem for finitely generated abelian groups*: any finitely generated abelian group  $A$  is the direct product of finitely many finite or infinite cyclic subgroups. The finite cyclic subgroups have orders  $d_1, \dots, d_s$ , where  $d_1 > 1$  if  $s > 0$  and  $d_i | d_j$  if  $i \leq j$ . In particular,  $A$  is the direct product of a finite subgroup  $A'$  (of order  $d_1 \cdots d_r$ ), its *torsion subgroup*, and a *free* subgroup  $A''$ .

The fundamental structure theorem also has an important application to linear algebra. Let  $V$  be a vector space over a field  $K$  and  $T : V \rightarrow V$  a linear transformation. We can give  $V$  the structure of a  $K[t]$ -module by defining, for any  $v \in V$  and any  $f = a_0 + a_1 t + \dots + a_n t^n \in K[t]$ ,

$$fv = a_0 v + a_1 T v + \dots + a_n T^n v.$$

If  $V$  is finite-dimensional, then for any  $v \in V$  there is a nonzero polynomial  $f$  such that  $fv = 0$ . In this case the fundamental structure theorem says that  $V$  is the direct sum of finitely many subspaces  $V_1, \dots, V_s$  which are invariant under  $T$ . If  $V_i$  has dimension  $n_i \geq 1$ , then there exists a vector  $w_i \in V_i$  such that  $w_i, T w_i, \dots, T^{n_i-1} w_i$  are a vector space basis for  $V_i$  ( $1 \leq i \leq s$ ). There is a uniquely determined monic polynomial  $m_i$  of degree  $n_i$  such that  $m_i(T)w_i = 0$  and, finally,  $m_i | m_j$  if  $i \leq j$ .

The Smith normal form can be used to solve systems of linear ordinary differential equations with constant coefficients. Such a system has the form

$$\begin{aligned}
a_{11}(D)x_1 + \cdots + a_{1n}(D)x_n &= c_1(t) \\
a_{21}(D)x_1 + \cdots + a_{2n}(D)x_n &= c_2(t) \\
&\vdots \\
a_{m1}(D)x_1 + \cdots + a_{mn}(D)x_n &= c_m(t),
\end{aligned}$$

where the coefficients  $a_{jk}(D)$  are polynomials in  $D = d/dt$  with complex coefficients and the right sides  $c_j(t)$  are, say, infinitely differentiable functions of the time  $t$ . Since  $\mathbb{C}[s]$  is a Euclidean domain, we can bring the coefficient matrix  $A = (a_{jk}(D))$  to Smith normal form and thus replace the given system by an equivalent system in which the variables are ‘uncoupled’.

For the polynomial ring  $R = K[t]$  it is possible to say more about  $R$ -modules than for an arbitrary Euclidean domain, since the absolute value

$$|f| = 2^{\partial(f)} \quad \text{if } f \neq 0, |0| = 0,$$

has not only the Euclidean property, but also the properties

$$|f + g| \leq \max\{|f|, |g|\}, \quad |fg| = |f||g| \quad \text{for any } f, g \in R.$$

For any  $\mathbf{a} \in R^m$ , where  $R = K[t]$ , define  $|\mathbf{a}|$  to be the maximum absolute value of any of its coordinates. Then a basis for a module  $\mathbf{M} \subseteq R^m$  can be obtained in the following way. Suppose  $\mathbf{M} \neq \mathbf{0}$  and choose a nonzero element  $\mathbf{a}_1$  of  $\mathbf{M}$  for which  $|\mathbf{a}_1|$  is a minimum. If there is an element of  $\mathbf{M}$  which is not of the form  $p_1\mathbf{a}_1$  with  $p_1 \in R$ , choose one,  $\mathbf{a}_2$ , for which  $|\mathbf{a}_2|$  is a minimum. If there is an element of  $\mathbf{M}$  which is not of the form  $p_1\mathbf{a}_1 + p_2\mathbf{a}_2$  with  $p_1, p_2 \in R$ , choose one,  $\mathbf{a}_3$ , for which  $|\mathbf{a}_3|$  is a minimum. And so on.

Evidently  $|\mathbf{a}_1| \leq |\mathbf{a}_2| \leq \cdots$ . We will show that  $\mathbf{a}_1, \mathbf{a}_2, \dots$  are linearly independent for as long as the process can be continued, and thus ultimately a basis is obtained.

If this is not the case, then there exists a positive integer  $k \leq m$  such that  $\mathbf{a}_1, \dots, \mathbf{a}_k$  are linearly independent, but  $\mathbf{a}_1, \dots, \mathbf{a}_{k+1}$  are not. Hence there exist  $s_1, \dots, s_{k+1} \in R$  with  $s_{k+1} \neq 0$  such that  $s_1\mathbf{a}_1 + \cdots + s_{k+1}\mathbf{a}_{k+1} = \mathbf{0}$ . For each  $j \leq k$ , there exist  $q_j, r_j \in R$  such that

$$s_j = q_js_{k+1} + r_j, \quad |r_j| < |s_{k+1}|.$$

Put

$$\mathbf{a}'_{k+1} = \mathbf{a}_{k+1} + q_1\mathbf{a}_1 + \cdots + q_k\mathbf{a}_k, \quad \mathbf{b}_k = r_1\mathbf{a}_1 + \cdots + r_k\mathbf{a}_k.$$

Since  $\mathbf{a}_{k+1}$  is not of the form  $p_1\mathbf{a}_1 + \cdots + p_k\mathbf{a}_k$ , neither is  $\mathbf{a}'_{k+1}$  and hence  $|\mathbf{a}'_{k+1}| \geq |\mathbf{a}_{k+1}|$ . Furthermore,  $|\mathbf{b}_k| \leq \max_{1 \leq j \leq k} |r_j||\mathbf{a}_j| < |s_{k+1}||\mathbf{a}_{k+1}|$ . Since  $\mathbf{b}_k = -s_{k+1}\mathbf{a}'_{k+1}$ , by construction, this is a contradiction.

A basis for  $\mathbf{M}$  which is obtained in this way will be called a *minimal basis*. It is not difficult to show that a basis  $\mathbf{a}_1, \dots, \mathbf{a}_n$  is a minimal basis if and only if  $|\mathbf{a}_1| \leq \cdots \leq |\mathbf{a}_n|$  and the sum  $|\mathbf{a}_1| + \cdots + |\mathbf{a}_n|$  is minimal. Although a minimal basis is not uniquely determined, the values  $|\mathbf{a}_1|, \dots, |\mathbf{a}_n|$  are uniquely determined.

## 5 Further Remarks

For the history of the law of quadratic reciprocity, see Frei [16]. The first two proofs by Gauss of the law of quadratic reciprocity appeared in §§125–145 and §262 of [17]. A simplified account of Gauss's inductive proof has been given by Brown [7]. The proofs most commonly given use 'Gauss's lemma' and are variants of Gauss's third proof. The first proof given here, due to Rousseau [46], is of this general type, but it does not use Gauss's lemma and is based on a natural definition of the Jacobi symbol. For an extension of this definition of Zolotareff to algebraic number fields, see Cartier [9].

For Dirichlet's evaluation of Gauss sums, see [33]. A survey of Gauss sums is given in Berndt and Evans [6].

The extension of the law of quadratic reciprocity to arbitrary algebraic number fields was the subject of Hilbert's 9th Paris problem. Although such generalizations lie outside the scope of the present work, it may be worthwhile to give a brief glimpse. Let  $K = \mathbb{Q}$  be the field of rational numbers and let  $L = \mathbb{Q}(\sqrt{d})$  be a quadratic extension of  $K$ . If  $p$  is a prime in  $K$ , the law of quadratic reciprocity may be interpreted as describing how the ideal generated by  $p$  in  $L$  factors into prime ideals. Now let  $K$  be an arbitrary algebraic number field and let  $L$  be any finite extension of  $K$ . Quite generally, we may ask how the arithmetic of the extension  $L$  is determined by the arithmetic of  $K$ . The general reciprocity law, conjectured by Artin in 1923 and proved by him in 1927, gives an answer in the form of an isomorphism between two groups, provided the Galois group of  $L$  over  $K$  is abelian. For an introduction, see Wyman [54] and, for more detail, Tate [51]. The outstanding problem is to find a meaningful extension to the case when the Galois group is non-abelian. Some intriguing conjectures are provided by the Langlands program, for which see also Gelbart [18].

The law of quadratic reciprocity has an analogue for polynomials with coefficients from a finite field. Let  $\mathbb{F}_q$  be a finite field containing  $q$  elements, where  $q$  is a power of an odd prime. If  $g \in \mathbb{F}_q[x]$  is a monic irreducible polynomial of positive degree, then for any  $f \in \mathbb{F}_q[x]$  not divisible by  $g$  we define  $(f/g)$  to be 1 if  $f$  is congruent to a square mod  $g$ , and  $-1$  otherwise. The law of quadratic reciprocity, which in the case of prime  $q$  was stated by Dedekind (1857) and proved by Artin (1924), says that

$$(f/g)(g/f) = (-1)^{mn(q-1)/2}$$

for any distinct monic irreducible polynomials  $f, g \in \mathbb{F}_q[x]$  of positive degrees  $m, n$ . Artin also developed a theory of ideals, analogous to that for quadratic number fields, for the field obtained by adjoining to  $\mathbb{F}_q[x]$  an element  $\omega$  with  $\omega^2 = D(x)$ , where  $D(x) \in \mathbb{F}_q[x]$  is square-free; see [3].

Quadratic fields are treated in the third volume of Landau [30]. There is also a useful resumé accompanying the tables in Ince [23].

A complex number is said to be *algebraic* if it is a root of a monic polynomial with rational coefficients and *transcendental* otherwise. Hence a complex number is algebraic if and only if it is an element of some algebraic number field.

For an introduction to the theory of algebraic number fields, see Samuel [47]. This vast theory may be approached in a variety of ways. For a more detailed treatment the student may choose from Hecke [22], Hasse [20], Lang [32], Narkiewicz [38] and Neukirch [39]. There are useful articles in Cassels and Fröhlich [10], and Artin [2] treats also algebraic functions.

For the early history of Fermat's last theorem, see Vandiver [52], Ribenboim [41] and Kummer [28]. Further references will be given in Chapter XIII.

Arithmetical functions are discussed in Apostol [1], McCarthy [35] and Sivaramakrishnan [48]. The term 'Dirichlet product' comes from the connection with Dirichlet series, which will be considered in Chapter IX, §6. The ring of all arithmetical functions was shown to be a factorial domain by Cashwell and Everett (1959); the result is proved in [48].

In the form  $f(a \wedge b)f(a \vee b) = f(a)f(b)$ , the concept of multiplicative function can be extended to any map  $f : L \rightarrow \mathbb{C}$ , where  $L$  is a lattice. Möbius inversion can be extended to any locally finite partially ordered set and plays a significant role in modern combinatorics; see Bender and Goldman [5], Rota [45] and Barnabei *et al.* [4].

The early history of perfect numbers and Fermat numbers is described in Dickson [13]. It has been proved that any odd perfect number, if such a thing exists, must be greater than  $10^{300}$  and have at least 8 distinct prime factors. On the other hand, if an odd perfect number  $N$  has at most  $k$  distinct prime factors, then  $N < 4^{4^k}$  and thus all such  $N$  can be found by a finite amount of computation. See de Riele [42] and Heath-Brown [21].

The proof of the Lucas–Lehmer test for Mersenne primes follows Rosen [43] and Bruce [8]. For the conjectured distribution of Mersenne primes, see Wagstaff [53]. The construction of regular polygons by ruler and compass is discussed in Hadlock [19], Jacobson [24] and Morandi [36].

Much of the material in §4 is also discussed in Macduffee [34] and Newman [40]. Corollary 34 was proved by Hermite (1849), who later (1851) also proved Corollary 38. Indeed the latter result is the essential content of *Hermite's normal form*, which will be encountered in Chapter VIII, §2.

It is clear that Corollary 34 remains valid if the underlying ring  $\mathbb{Z}$  is replaced by any principal ideal domain. There have recently been some noteworthy extensions to more general rings. It may be asked, for an arbitrary commutative ring  $R$  and any  $a_1, \dots, a_n \in R$ , does there exist an invertible  $n \times n$  matrix  $U$  with entries from  $R$  which has  $a_1, \dots, a_n$  as its first row? It is obviously necessary that there exist  $x_1, \dots, x_n \in R$  such that

$$a_1x_1 + \dots + a_nx_n = 1,$$

i.e. that the ideal generated by  $a_1, \dots, a_n$  be the whole ring  $R$ . If  $n = 2$ , this necessary condition is also sufficient, by the same observation as when invertibility of matrices was first considered for  $R = \mathbb{Z}$ . However, if  $n > 2$  there exist even factorial domains  $R$  for which the condition is not sufficient. In 1976 Quillen and Suslin independently proved the twenty-year-old conjecture that it is sufficient if  $R = K[t_1, \dots, t_d]$  is the ring of polynomials in finitely many indeterminates with coefficients from an arbitrary field  $K$ .

By pursuing an analogy between projective modules in algebra and vector bundles in topology, Serre (1955) had been led to conjecture that, for  $R = K[t_1, \dots, t_d]$ , if an  $R$ -module has a finite basis and is the direct sum of two submodules, then each of these submodules has a finite basis. Seshadri (1958) proved the conjecture for  $d = 2$  and in the same year Serre showed that, for arbitrary  $d$ , it would follow from the result which Quillen and Suslin subsequently proved.

For proofs of these results and for later developments, see Lam [29], Fitchas and Galligo [14], and Swan [50]. There is a short proof of the Quillen–Suslin theorem in Lang [31].

For Smith’s normal form, see Smith [49] and Kaplansky [27]. It was shown by Wedderburn (1915) that Smith’s normal form also holds for matrices of holomorphic functions, even though the latter do not form a principal ideal domain; see Narasimhan [37].

Finitely generated commutative groups are important, not only because more can be said about them, but also because they arise in practice. *Dirichlet’s unit theorem* says that the units of an algebraic number field form a finitely generated commutative group. As will be seen in Chapter XIII, §4, *Mordell’s theorem* says that the rational points of an elliptic curve also form a finitely generated commutative group.

Modules over a polynomial ring  $K[s]$  play an important role in what electrical engineers call *linear systems theory*. Connected accounts are given in Kalman [26], Rosenbrock [44] and Kailath [25]. For some further mathematical developments, see Forney [15], Coppel [11], and Coppel and Cullen [12].

## 6 Selected References

- [1] T.M. Apostol, *Introduction to analytic number theory*, Springer-Verlag, New York, 1976.
- [2] E. Artin, *Algebraic numbers and algebraic functions*, Nelson, London, 1968.
- [3] E. Artin, Quadratische Körper im Gebiet der höheren Kongruenzen I, II, *Collected Papers* (ed. S. Lang and J.T. Tate), pp. 1–94, reprinted, Springer-Verlag, New York, 1986.
- [4] M. Barnabei, A. Brini and G.-C. Rota, The theory of Möbius functions, *Russian Math. Surveys* **41** (1986), no. 3, 135–188.
- [5] E.A. Bender and J.R. Goldman, On the application of Möbius inversion in combinatorial analysis, *Amer. Math. Monthly* **82** (1975), 789–803.
- [6] B.C. Berndt and R.J. Evans, The determination of Gauss sums, *Bull. Amer. Math. Soc. (N.S.)* **5** (1981), 107–129.
- [7] E. Brown, The first proof of the quadratic reciprocity law, revisited, *Amer. Math. Monthly* **88** (1981), 257–264.
- [8] J.W. Bruce, A really trivial proof of the Lucas–Lehmer test, *Amer. Math. Monthly* **100** (1993), 370–371.
- [9] P. Cartier, Sur une généralisation des symboles de Legendre–Jacobi, *Enseign. Math.* **16** (1970), 31–48.
- [10] J.W.S. Cassels and A. Fröhlich (ed.), *Algebraic number theory*, Academic Press, London, 1967.
- [11] W.A. Coppel, Matrices of rational functions, *Bull. Austral. Math. Soc.* **11** (1974), 89–113.
- [12] W.A. Coppel and D.J. Cullen, Strong system equivalence (II), *J. Austral. Math. Soc. B* **27** (1985), 223–237.
- [13] L.E. Dickson, *History of the theory of numbers, Vol. I*, reprinted, Chelsea, New York, 1992.
- [14] N. Fitchas and A. Galligo, Nullstellensatz effectif et conjecture de Serre (théorème de Quillen–Suslin) pour le calcul formel, *Math. Nachr.* **149** (1990), 231–253.
- [15] G.D. Forney, Minimal bases of rational vector spaces, with applications to multivariable linear systems, *SIAM J. Control* **13** (1975), 493–520.
- [16] G. Frei, The reciprocity law from Euler to Eisenstein, *The intersection of history and mathematics* (ed. C. Sasaki, M. Sugiura and J.W. Dauben), pp. 67–90, Birkhäuser, Basel, 1994.
- [17] C.F. Gauss, *Disquisitiones arithmeticae*, English translation by A.A. Clarke, revised by W.C. Waterhouse, Springer, New York, 1986. [Latin original, 1801]

- [18] S. Gelbart, An elementary introduction to the Langlands program, *Bull. Amer. Math. Soc. (N.S.)* **10** (1984), 177–219.
- [19] C.R. Hadlock, *Field theory and its classical problems*, Carus Mathematical Monographs no. 19, Mathematical Association of America, Washington, D.C., 1978. [Reprinted in paperback, 2000]
- [20] H. Hasse, *Number theory*, English transl. by H.G. Zimmer, Springer-Verlag, Berlin, 1980.
- [21] D.R. Heath-Brown, Odd perfect numbers, *Math. Proc. Cambridge Philos. Soc.* **115** (1994), 191–196.
- [22] E. Hecke, *Lectures on the theory of algebraic numbers*, English transl. by G.U. Brauer, J.R. Goldman and R. Kotzen, Springer-Verlag, New York, 1981. [German original, 1923]
- [23] E.L. Ince, *Cycles of reduced ideals in quadratic fields*, Mathematical Tables Vol. IV, British Association, London, 1934.
- [24] N. Jacobson, *Basic Algebra I*, 2nd ed., W.H. Freeman, New York, 1985.
- [25] T. Kailath, *Linear systems*, Prentice–Hall, Englewood Cliffs, N.J., 1980.
- [26] R.E. Kalman, Algebraic theory of linear systems, *Topics in mathematical system theory* (R.E. Kalman, P.L. Falb and M.A. Arbib), pp. 237–339, McGraw–Hill, New York, 1969.
- [27] I. Kaplansky, Elementary divisors and modules, *Trans. Amer. Math. Soc.* **66** (1949), 464–491.
- [28] E. Kummer, *Collected Papers, Vol. I* (ed. A. Weil), Springer-Verlag, Berlin, 1975.
- [29] T.Y. Lam, *Serre’s conjecture*, Lecture Notes in Mathematics **635**, Springer-Verlag, Berlin, 1978.
- [30] E. Landau, *Vorlesungen über Zahlentheorie*, 3 vols., Hirzel, Leipzig, 1927. [Reprinted, Chelsea, New York, 1969]
- [31] S. Lang, *Algebra*, corrected reprint of 3rd ed., Addison-Wesley, Reading, Mass., 1994.
- [32] S. Lang, *Algebraic number theory*, 2nd ed., Springer-Verlag, New York, 1994.
- [33] G. Lejeune-Dirichlet, *Werke*, Band I, pp. 237–256, reprinted Chelsea, New York, 1969.
- [34] C.C. Macduffee, *The theory of matrices*, corrected reprint, Chelsea, New York, 1956.
- [35] P.J. McCarthy, *Introduction to arithmetical functions*, Springer-Verlag, New York, 1986.
- [36] P. Morandi, *Field and Galois theory*, Springer-Verlag, New York, 1996.
- [37] R. Narasimhan, *Complex analysis in one variable*, Birkhäuser, Boston, Mass., 1985.
- [38] W. Narkiewicz, *Elementary and analytic theory of algebraic numbers*, 2nd ed., Springer-Verlag, Berlin, 1990.
- [39] J. Neukirch, *Algebraic number theory*, English transl. by N. Schappacher, Springer, Berlin, 1999.
- [40] M. Newman, *Integral matrices*, Academic Press, New York, 1972.
- [41] P. Ribenboim, *13 Lectures on Fermat’s last theorem*, Springer-Verlag, New York, 1979.
- [42] H.J.J. te Riele, Perfect numbers and aliquot sequences, *Computational methods in number theory* (ed. H.W. Lenstra Jr. and R. Tijdeman), Part I, pp. 141–157, Mathematical Centre Tracts **154**, Amsterdam, 1982.
- [43] M.L. Rosen, A proof of the Lucas–Lehmer test, *Amer. Math. Monthly* **95** (1988), 855–856.
- [44] H.H. Rosenbrock, *State-space and multivariable theory*, Nelson, London, 1970.
- [45] G.-C. Rota, On the foundations of combinatorial theory I. Theory of Möbius functions, *Z. Wahrsch. Verw. Gebiete* **2** (1964), 340–368.
- [46] G. Rousseau, On the Jacobi symbol, *J. Number Theory* **48** (1994), 109–111.
- [47] P. Samuel, *Algebraic theory of numbers*, English transl. by A.J. Silberger, Houghton Mifflin, Boston, Mass., 1970.
- [48] R. Sivaramakrishnan, *Classical theory of arithmetic functions*, M. Dekker, New York, 1989.
- [49] H.J.S. Smith, *Collected mathematical papers, Vol. 1*, pp. 367–409, reprinted, Chelsea, New York, 1965.

- [50] R.G. Swan, Gubeladze's proof of Anderson's conjecture, *Azumaya algebras, actions and modules* (ed. D. Haile and J. Osterburg), pp. 215–250, Contemporary Mathematics **124**, Amer. Math. Soc., Providence, R.I., 1992.
- [51] J. Tate, Problem 9: The general reciprocity law, *Mathematical developments arising from Hilbert problems* (ed. F.E. Browder), pp. 311–322, Proc. Symp. Pure Math. **28**, Part 2, Amer. Math. Soc., Providence, Rhode Island, 1976.
- [52] H.S. Vandiver, Fermat's last theorem: its history and the nature of the known results concerning it, *Amer. Math. Monthly* **53** (1946), 555–578.
- [53] S.S. Wagstaff Jr., Divisors of Mersenne numbers, *Math. Comp.* **40** (1983), 385–397.
- [54] B.F. Wyman, What is a reciprocity law?, *Amer. Math. Monthly* **79** (1972), 571–586.

## Additional References

- J. Bernstein and S. Gelbart (ed.), *Introduction to the Langlands program*, Birkhäuser, Boston, 2003.
- E. Frenkel, Recent advances in the Langlands program, *Bull. Amer. Math. Soc. (N.S.)* **41** (2004), 151–184.
- T. Metsänkylä, Catalan's conjecture: another old Diophantine problem solved, *Bull. Amer. Math. Soc. (N.S.)* **41** (2004), 43–57.

## IV

### Continued Fractions and Their Uses

#### 1 The Continued Fraction Algorithm

Let  $\xi = \xi_0$  be an irrational real number. Then we can write

$$\xi_0 = a_0 + \xi_1^{-1},$$

where  $a_0 = \lfloor \xi_0 \rfloor$  is the greatest integer  $\leq \xi_0$  and where  $\xi_1 > 1$  is again an irrational number. Hence the process can be repeated indefinitely:

$$\begin{aligned}\xi_1 &= a_1 + \xi_2^{-1}, & (a_1 = \lfloor \xi_1 \rfloor, \xi_2 > 1), \\ \xi_2 &= a_2 + \xi_3^{-1}, & (a_2 = \lfloor \xi_2 \rfloor, \xi_3 > 1), \\ &\dots\end{aligned}$$

By construction,  $a_n \in \mathbb{Z}$  for all  $n \geq 0$  and  $a_n \geq 1$  if  $n \geq 1$ . The uniquely determined infinite sequence  $[a_0, a_1, a_2, \dots]$  is called the *continued fraction expansion* of  $\xi$ . The continued fraction expansion of  $\xi_n$  is  $[a_n, a_{n+1}, a_{n+2}, \dots]$ .

For example, the ‘golden ratio’  $\tau = (1 + \sqrt{5})/2$  has the continued fraction expansion  $[1, 1, 1, \dots]$ , since  $\tau = 1 + \tau^{-1}$ . Similarly,  $\sqrt{2}$  has the continued fraction expansion  $[1, 2, 2, \dots]$ , since  $\sqrt{2} + 1 = 2 + 1/(\sqrt{2} + 1)$ .

The relation between  $\xi_n$  and  $\xi_{n+1}$  may be written as a linear fractional transformation:

$$\xi_n = (a_n \xi_{n+1} + 1)/(1 \xi_{n+1} + 0).$$

An arbitrary linear fractional transformation

$$\xi = (\alpha \xi' + \beta)/(\gamma \xi' + \delta)$$

is completely determined by its matrix

$$T = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}.$$

This description is convenient, because if we make a further linear fractional transformation

$$\xi' = (\alpha' \xi'' + \beta')/(\gamma' \xi'' + \delta')$$

with matrix

$$T' = \begin{pmatrix} \alpha' & \beta' \\ \gamma' & \delta' \end{pmatrix},$$

then, as is easily verified, the matrix

$$T'' = \begin{pmatrix} \alpha'' & \beta'' \\ \gamma'' & \delta'' \end{pmatrix}$$

of the composite transformation

$$\xi = (\alpha''\xi'' + \beta'')/(\gamma''\xi'' + \delta'')$$

is given by the matrix product  $T'' = TT'$ .

It follows that, if we set

$$A_k = \begin{pmatrix} a_k & 1 \\ 1 & 0 \end{pmatrix},$$

then the matrix of the linear fractional transformation which expresses  $\xi$  in terms of  $\xi_{n+1}$  is

$$T_n = A_0 \cdots A_n.$$

It is readily verified by induction that

$$T_n = \begin{pmatrix} p_n & p_{n-1} \\ q_n & q_{n-1} \end{pmatrix},$$

i.e.,

$$\xi = (p_n \xi_{n+1} + p_{n-1})/(q_n \xi_{n+1} + q_{n-1}),$$

where the elements  $p_n, q_n$  satisfy the recurrence relations

$$p_n = a_n p_{n-1} + p_{n-2}, \quad q_n = a_n q_{n-1} + q_{n-2} \quad (n \geq 0), \quad (1)$$

with the conventional starting values

$$p_{-2} = 0, \quad p_{-1} = 1, \quad \text{resp. } q_{-2} = 1, \quad q_{-1} = 0. \quad (2)$$

In particular,

$$p_0 = a_0, \quad p_1 = a_1 a_0 + 1, \quad q_0 = 1, \quad q_1 = a_1.$$

Since  $\det A_k = -1$ , by taking determinants we obtain

$$p_n q_{n-1} - p_{n-1} q_n = (-1)^{n+1} \quad (n \geq 0). \quad (3)$$

By (1) also,

$$\begin{pmatrix} p_n & p_{n-1} \\ q_n & q_{n-1} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -a_n & 0 \end{pmatrix} = \begin{pmatrix} p_{n-2} & p_n \\ q_{n-2} & q_n \end{pmatrix},$$

from which, by taking determinants again, we get

$$p_n q_{n-2} - p_{n-2} q_n = (-1)^n a_n \quad (n \geq 0). \quad (4)$$

It follows from (1)–(2) that  $p_n$  and  $q_n$  are integers, and from (3) that they are coprime. Since  $a_n \geq 1$  for  $n \geq 1$ , we have

$$1 = q_0 \leq q_1 < q_2 < \dots$$

Thus  $q_n \geq n$  for  $n \geq 1$ . (In fact, since  $q_n \geq q_{n-1} + q_{n-2}$  for  $n \geq 1$ , it is readily shown by induction that  $q_n > \tau^{n-1}$  for  $n > 1$ , where  $\tau = (1 + \sqrt{5})/2$ .)

Since  $q_n > 0$  for  $n \geq 0$ , we can rewrite (3), (4) in the forms

$$p_n/q_n - p_{n-1}/q_{n-1} = (-1)^{n+1}/q_{n-1}q_n \quad (n \geq 1), \quad (3)'$$

$$p_n/q_n - p_{n-2}/q_{n-2} = (-1)^n a_n/q_{n-2}q_n \quad (n \geq 2). \quad (4)'$$

It follows that the sequence  $\{p_{2n}/q_{2n}\}$  is increasing, the sequence  $\{p_{2n+1}/q_{2n+1}\}$  is decreasing, and every member of the first sequence is less than every member of the second sequence. Hence both sequences have limits and actually, since  $q_n \rightarrow \infty$ , the limits of the two sequences are the same.

From

$$\xi = (p_n \xi_{n+1} + p_{n-1})/(q_n \xi_{n+1} + q_{n-1})$$

we obtain

$$q_n \xi - p_n = (p_{n-1} q_n - p_n q_{n-1})/(q_n \xi_{n+1} + q_{n-1}) = (-1)^n/(q_n \xi_{n+1} + q_{n-1}).$$

Hence  $\xi > p_n/q_n$  if  $n$  is even and  $\xi < p_n/q_n$  if  $n$  is odd. It follows that  $p_n/q_n \rightarrow \xi$  as  $n \rightarrow \infty$ . Consequently different irrational numbers have different continued fraction expansions.

Since  $\xi$  lies between  $p_n/q_n$  and  $p_{n+1}/q_{n+1}$ , we have

$$|p_{n+2}/q_{n+2} - p_n/q_n| < |\xi - p_n/q_n| < |p_{n+1}/q_{n+1} - p_n/q_n|.$$

By (3)' and (4)' we can rewrite this in the form

$$a_{n+2}/q_n q_{n+2} < |\xi - p_n/q_n| < 1/q_n q_{n+1} \quad (n \geq 0). \quad (5)$$

Hence

$$q_{n+2}^{-1} < |q_n \xi - p_n| < q_{n+1}^{-1},$$

which shows that  $|q_n \xi - p_n|$  decreases as  $n$  increases. It follows that  $|\xi - p_n/q_n|$  also decreases as  $n$  increases.

The rational number  $p_n/q_n$  is called the  $n$ -th *convergent* of  $\xi$ . The integers  $a_n$  are called the *partial quotients* and the real numbers  $\xi_n$  the *complete quotients* in the continued fraction expansion of  $\xi$ .

The continued fraction algorithm can be applied also when  $\xi = \xi_0$  is rational, but in this case it is really the same as the Euclidean algorithm. For suppose  $\xi_n = b_n/c_n$ .

where  $b_n$  and  $c_n$  are integers and  $c_n > 0$ . We can write

$$b_n = a_n c_n + c_{n+1},$$

where  $a_n = \lfloor \xi_n \rfloor$  and  $c_{n+1}$  is an integer such that  $0 \leq c_{n+1} < c_n$ . Thus  $\xi_{n+1}$  is defined if and only if  $c_{n+1} \neq 0$ , and then  $\xi_{n+1} = c_n / c_{n+1}$ . Since the sequence of positive integers  $\{c_n\}$  cannot decrease for ever, the continued fraction algorithm for a rational number  $\xi$  always terminates. At the last stage of the algorithm we have simply

$$\xi_N = a_N,$$

where  $a_N > 1$  if  $N > 0$ . The uniquely determined finite sequence  $[a_0, a_1, \dots, a_N]$  is called the continued fraction expansion of  $\xi$ .

*Convergents* and *complete quotients* can be defined as before; all the properties derived for  $\xi$  irrational continue to hold for  $\xi$  rational, provided we do not go past  $n = N$ . The relation

$$\xi = (p_{N-1}\xi_N + p_{N-2}) / (q_{N-1}\xi_N + q_{N-2})$$

now shows that

$$\xi = p_N / q_N.$$

Consequently different rational numbers have different continued fraction expansions.

Now let  $a_0, a_1, a_2, \dots$  be any infinite sequence of integers with  $a_n \geq 1$  for  $n \geq 1$ . If we define integers  $p_n, q_n$  by the recurrence relations (1)–(2), our previous argument shows that the sequence  $\{p_{2n}/q_{2n}\}$  is increasing, the sequence  $\{p_{2n+1}/q_{2n+1}\}$  is decreasing, and the two sequences have a common limit  $\xi$ . If we put  $\xi_0 = \xi$  and

$$\xi_{n+1} = -(q_{n-1}\xi - p_{n-1}) / (q_n\xi - p_n) \quad (n \geq 0),$$

our previous argument shows also that  $\xi_{n+1} > 1$  ( $n \geq 0$ ). Since

$$\xi_n = a_n + \xi_{n+1}^{-1},$$

it follows that  $a_n = \lfloor \xi_n \rfloor$ . Hence  $\xi$  is irrational and  $[a_0, a_1, a_2, \dots]$  is its continued fraction expansion.

Similarly it may be seen that, for any finite sequence of integers  $a_0, a_1, \dots, a_N$ , with  $a_n \geq 1$  for  $1 \leq n < N$  and  $a_N > 1$  if  $N > 0$ , there is a rational number  $\xi$  with  $[a_0, a_1, \dots, a_N]$  as its continued fraction expansion.

We will write simply  $\xi = [a_0, a_1, \dots, a_N]$  if  $\xi$  is rational and  $\xi = [a_0, a_1, a_2, \dots]$  if  $\xi$  is irrational.

We will later have need of the following result:

**Lemma 0** *Let  $\xi$  be an irrational number with complete quotients  $\xi_n$  and convergents  $p_n/q_n$ . If  $\eta$  is any irrational number different from  $\xi$ , and if we define  $\eta_{n+1}$  by*

$$\eta = (p_n\eta_{n+1} + p_{n-1}) / (q_n\eta_{n+1} + q_{n-1}),$$

*then  $-1 < \eta_n < 0$  for all large  $n$ .*

*Moreover, if  $\xi > 1$  and  $\eta < 0$ , then  $-1 < \eta_n < 0$  for all  $n > 0$ .*

*Proof* We have

$$\eta_{n+1} = (q_{n-1}\eta - p_{n-1})/(p_n - q_n\eta).$$

Hence

$$\begin{aligned}\theta_{n+1} &:= q_n\eta_{n+1} + q_{n-1} \\ &= (p_nq_{n-1} - p_{n-1}q_n)/(p_n - q_n\eta) \\ &= (-1)^{n+1}/(p_n - q_n\eta) \\ &= (-1)^n/q_n(\eta - p_n/q_n).\end{aligned}$$

Since  $p_n/q_n \rightarrow \xi \neq \eta$  and  $q_n \rightarrow \infty$ , it follows that  $\theta_n \rightarrow 0$ . Since

$$\eta_{n+1} = -(q_{n-1} - \theta_{n+1})/q_n,$$

we conclude that  $-1 < \eta_{n+1} < 0$  for all large  $n$ .

Suppose now that  $\xi > 1$  and  $\eta < 0$ . It is readily verified that  $\eta_n = a_n + 1/\eta_{n+1}$ . But  $a_n = \lfloor \xi_n \rfloor \geq 1$  for all  $n \geq 0$ . Consequently  $\eta_n < 0$  implies  $1/\eta_{n+1} < -1$  and thus  $-1 < \eta_{n+1} < 0$ . Since  $\eta_0 < 0$ , it follows by induction that  $-1 < \eta_n < 0$  for all  $n > 0$ .  $\square$

The complete quotients of a real number may be characterized in the following way:

**Proposition 1** *If  $\eta > 1$  and*

$$\xi = (p\eta + p')/(q\eta + q'),$$

*where  $p, q, p', q'$  are integers such that  $pq' - p'q = \pm 1$  and  $q > q' > 0$ , then  $\eta$  is a complete quotient of  $\xi$  and  $p'/q', p/q$  are corresponding consecutive convergents of  $\xi$ .*

*Proof* The relation  $pq' - p'q = \pm 1$  implies that  $p$  and  $q$  are relatively prime. Since  $q > 0$ ,  $p/q$  has a finite continued fraction expansion

$$p/q = [a_0, a_1, \dots, a_{n-1}] = p_{n-1}/q_{n-1}$$

and  $q = q_{n-1}$ ,  $p = p_{n-1}$ . In fact, since  $q > 1$ , we have  $n > 1$ ,  $a_{n-1} \geq 2$  and  $q_{n-1} > q_{n-2}$ . From

$$p_{n-1}q_{n-2} - p_{n-2}q_{n-1} = (-1)^n = \varepsilon(pq' - p'q),$$

where  $\varepsilon = \pm 1$ , we obtain

$$p_{n-1}(q_{n-2} - \varepsilon q') = q_{n-1}(p_{n-2} - \varepsilon p').$$

Hence  $q_{n-1}$  divides  $q_{n-2} - \varepsilon q'$ . Since  $0 < q_{n-2} < q_{n-1}$  and  $0 < q' < q_{n-1}$ , it follows that  $q' = q_{n-2}$  if  $\varepsilon = 1$  and  $q' = q_{n-1} - q_{n-2}$  if  $\varepsilon = -1$ . Hence  $p' = p_{n-2}$  if  $\varepsilon = 1$  and  $p' = p_{n-1} - p_{n-2}$  if  $\varepsilon = -1$ . Thus

$$\begin{aligned}\xi &= (p_{n-1}\eta + p_{n-2})/(q_{n-1}\eta + q_{n-2}), \\ \text{resp. } &(p_{n-1}\eta + p_{n-1} - p_{n-2})/(q_{n-1}\eta + q_{n-1} - q_{n-2}).\end{aligned}$$

Since  $\eta > 1$ , its continued fraction expansion has the form  $[a_n, a_{n+1}, \dots]$ , where  $a_n \geq 1$ . It follows that  $\zeta$  has the continued fraction expansion

$$[a_0, a_1, \dots, a_{n-1}, a_n, \dots], \text{ resp. } [a_0, a_1, \dots, a_{n-1} - 1, 1, a_n, \dots].$$

In either case  $p'/q'$  and  $p/q$  are consecutive convergents of  $\zeta$  and  $\eta$  is the corresponding complete quotient.  $\square$

A complex number  $\zeta$  is said to be *equivalent* to a complex number  $\omega$  if there exist integers  $a, b, c, d$  with  $ad - bc = \pm 1$  such that

$$\zeta = (a\omega + b)/(c\omega + d),$$

and *properly equivalent* if actually  $ad - bc = 1$ . Then  $\omega$  is also equivalent, resp. properly equivalent, to  $\zeta$ , since

$$\omega = (d\zeta - b)/(-c\zeta + a).$$

By taking  $a = d = 1$  and  $b = c = 0$ , we see that any complex number  $\zeta$  is properly equivalent to itself. It is not difficult to verify also that if  $\zeta$  is equivalent to  $\omega$  and  $\omega$  equivalent to  $\chi$ , then  $\zeta$  is equivalent to  $\chi$ , and the same holds with ‘equivalence’ replaced by ‘proper equivalence’. Thus equivalence and proper equivalence are indeed ‘equivalence relations’.

For any coprime integers  $b, d$ , there exist integers  $a, c$  such that  $ad - bc = 1$ . Since

$$b/d = (a \cdot 0 + b)/(c \cdot 0 + d),$$

it follows that any rational number is properly equivalent to 0, and hence any two rational numbers are properly equivalent. The situation is more interesting for irrational numbers:

**Proposition 2** *Two irrational numbers  $\zeta, \eta$  are equivalent if and only if their continued fraction expansions  $[a_0, a_1, a_2, \dots]$ ,  $[b_0, b_1, b_2, \dots]$  have the same ‘tails’, i.e. there exist integers  $m \geq 0$  and  $n \geq 0$  such that*

$$a_{m+k} = b_{n+k} \quad \text{for all } k \geq 0.$$

*Proof* If the continued fraction expansions of  $\zeta$  and  $\eta$  have the same tails, then some complete quotient  $\zeta_m$  of  $\zeta$  coincides with some complete quotient  $\eta_n$  of  $\eta$ . But  $\zeta$  is equivalent to  $\zeta_m$ , since  $\zeta = (p_{m-1}\zeta_m + p_{m-2})/(q_{m-1}\zeta_m + q_{m-2})$  and  $p_{m-1}q_{m-2} - p_{m-2}q_{m-1} = (-1)^m$ , and similarly  $\eta$  is equivalent to  $\eta_n$ . Hence  $\zeta$  and  $\eta$  are equivalent.

Suppose on the other hand that  $\zeta$  and  $\eta$  are equivalent. Then

$$\eta = (a\zeta + b)/(c\zeta + d)$$

for some integers  $a, b, c, d$  such that  $ad - bc = \pm 1$ . By changing the signs of all four we may suppose that  $c\zeta + d > 0$ . From the relation

$$\zeta = (p_{n-1}\zeta_n + p_{n-2})/(q_{n-1}\zeta_n + q_{n-2})$$

between  $\xi$  and its complete quotient  $\xi_n$  it follows that

$$\eta = (a_n \xi_n + b_n) / (c_n \xi_n + d_n),$$

where

$$\begin{aligned} a_n &= ap_{n-1} + bq_{n-1}, & b_n &= ap_{n-2} + bq_{n-2}, \\ c_n &= cp_{n-1} + dq_{n-1}, & d_n &= cp_{n-2} + dq_{n-2}, \end{aligned}$$

and hence

$$a_n d_n - b_n c_n = (ad - bc)(p_{n-1} q_{n-2} - p_{n-2} q_{n-1}) = \pm 1.$$

The inequalities

$$|q_{n-1} \xi - p_{n-1}| < 1/q_n, \quad |q_{n-2} \xi - p_{n-2}| < 1/q_{n-1}$$

imply that

$$|c_n - (c\xi + d)q_{n-1}| < |c|/q_n, \quad |d_n - (c\xi + d)q_{n-2}| < |c|/q_{n-1}.$$

Since  $c\xi + d > 0$ ,  $q_{n-1} > q_{n-2}$  and  $q_n \rightarrow \infty$  as  $n \rightarrow \infty$ , it follows that  $c_n > d_n > 0$  for sufficiently large  $n$ . Then, by Proposition 1,  $\xi_n$  is a complete quotient also of  $\eta$ . Thus the continued fraction expansions of  $\xi$  and  $\eta$  have a common tail.  $\square$

## 2 Diophantine Approximation

The subject of *Diophantine approximation* is concerned with finding integer or rational solutions for systems of inequalities. For problems in one dimension the continued fraction algorithm is a most helpful tool, as we will now see.

**Proposition 3** *Let  $p_n/q_n$  ( $n \geq 1$ ) be a convergent of the real number  $\xi$ . If  $p, q$  are integers such that  $0 < q \leq q_n$  and  $p \neq p_n$  if  $q = q_n$ , then*

$$|q\xi - p| \geq |q_{n-1}\xi - p_{n-1}| > |q_n\xi - p_n|$$

and

$$|\xi - p/q| > |\xi - p_n/q_n|.$$

*Proof* It follows from (3) that the simultaneous linear equations

$$\lambda p_{n-1} + \mu p_n = p, \quad \lambda q_{n-1} + \mu q_n = q,$$

have integer solutions, namely

$$\lambda = (-1)^{n-1}(p_n q - q_n p), \quad \mu = (-1)^n(p_{n-1} q - q_{n-1} p).$$

The hypotheses on  $p, q$  imply that  $\lambda \neq 0$ . If  $\mu = 0$ , then

$$|q\xi - p| = |\lambda(q_{n-1}\xi - p_{n-1})| \geq |q_{n-1}\xi - p_{n-1}|.$$

Thus we now assume  $\mu \neq 0$ . Since  $q \leq q_n$ ,  $\lambda$  and  $\mu$  cannot both be positive and hence, since  $q > 0$ ,  $\lambda\mu < 0$ . Then

$$q\zeta - p = \lambda(q_{n-1}\zeta - p_{n-1}) + \mu(q_n\zeta - p_n)$$

and both terms on the right have the same sign. Hence

$$\begin{aligned} |q\zeta - p| &= |\lambda(q_{n-1}\zeta - p_{n-1})| + |\mu(q_n\zeta - p_n)| \\ &\geq |q_{n-1}\zeta - p_{n-1}|. \end{aligned}$$

This proves the first statement of the proposition. The second statement follows, since

$$\begin{aligned} |\zeta - p/q| &= q^{-1}|q\zeta - p| > q^{-1}|q_n\zeta - p_n| \\ &= (q_n/q)|\zeta - p_n/q_n| \\ &\geq |\zeta - p_n/q_n|. \end{aligned} \quad \square$$

To illustrate the application of Proposition 3, consider the continued fraction expansion of  $\pi = 3.14159265358 \dots$ . We easily find that it begins  $[3, 7, 15, 1, 292, \dots]$ . It follows that the first five convergents of  $\pi$  are

$$3/1, \quad 22/7, \quad 333/106, \quad 355/113, \quad 103993/33102.$$

Using the inequality  $|\zeta - p_n/q_n| < 1/q_n q_{n+1}$  and choosing  $n = 3$  so that  $q_{n+1}$  is large, we obtain

$$0 < 355/113 - \pi < 0.000000267 \dots$$

The approximation  $355/113$  to  $\pi$  was first given by the Chinese mathematician Zu Chongzhi in the 5th century A.D. Proposition 3 shows that it is a better approximation to  $\pi$  than any other rational number with denominator  $\leq 113$ .

In general, a rational number  $p'/q'$ , where  $p', q'$  are integers and  $q' > 0$ , may be said to be a *best approximation* to a real number  $\zeta$  if

$$|\zeta - p/q| > |\zeta - p'/q'|$$

for all different rational numbers  $p/q$  whose denominator  $q$  satisfies  $0 < q \leq q'$ . Thus Proposition 3 says that any convergent  $p_n/q_n$  ( $n \geq 1$ ) of  $\zeta$  is a best approximation of  $\zeta$ . However, these are not the only best approximations. It may be shown that, if  $p_{n-2}/q_{n-2}$  and  $p_{n-1}/q_{n-1}$  are consecutive convergents of  $\zeta$ , then any rational number of the form

$$(cp_{n-1} + p_{n-2})/(cq_{n-1} + q_{n-2}),$$

where  $c$  is an integer such that  $a_n/2 < c \leq a_n$  is a best approximation of  $\zeta$ . Furthermore, every best approximation of  $\zeta$  has this form if, when  $a_n$  is even, one allows also  $c = a_n/2$ .

It follows that  $355/113$  is a better approximation to  $\pi$  than any other rational number with denominator less than 16604, since  $292/2 = 146$  and  $146 \times 113 + 106 = 16604$ .

The complete continued fraction expansion of  $\pi$  is not known. However, it was discovered by Cotes (1714) and then proved by Euler (1737) that the complete continued fraction expansion of  $e = 2.71828182459\dots$  is given by  $e - 1 = [1, 1, 2, 1, 1, 4, 1, 1, 6, \dots]$ .

The preceding results may also be applied to the construction of calendars. The solar year has a length of about 365.24219 mean solar days. The continued fraction expansion of  $\lambda = (0.24219)^{-1}$  begins  $[4, 7, 1, 3, 24, \dots]$ . Hence the first five convergents of  $\lambda$  are

$$4/1, \quad 29/7, \quad 33/8, \quad 128/31, \quad 3105/752.$$

It follows that

$$0 < 128/31 - \lambda < 0.0000428$$

and 128/31 is a better approximation to  $\lambda$  than any other rational number with denominator less than 380. The Julian calendar, by adding a day every 4 years, estimated the year at 365.25 days. The Gregorian calendar, by adding 97 days every 400 years, estimates the year at 365.2425 days. Our analysis shows that, if we added instead 31 days every 128 years, we would obtain the much more precise estimate of 365.2421875 days.

Best approximations also find an application in the selection of gear ratios, and continued fractions were already used for this purpose by Huygens (1682) in constructing his planetarium (a mechanical model for the solar system).

The next proposition describes another way in which the continued fraction expansion provides good rational approximations.

**Proposition 4** *If  $p, q$  are coprime integers with  $q > 0$  such that, for some real number  $\xi$ ,*

$$|\xi - p/q| < 1/2q^2,$$

*then  $p/q$  is a convergent of  $\xi$ .*

*Proof* Let  $p_n/q_n$  be the convergents of  $\xi$  and assume that  $p/q$  is not a convergent. We show first that  $q < q_N$  for some  $N > 0$ . This is obvious if  $\xi$  is irrational. If  $\xi = p_N/q_N$  is rational, then

$$1/q_N \leq |qp_N - pq_N|/q_N = |q\xi - p| < 1/2q.$$

Hence  $q < q_N$  and  $N > 0$ .

It follows that  $q_{n-1} \leq q < q_n$  for some  $n > 0$ . By Proposition 3,

$$|q_{n-1}\xi - p_{n-1}| \leq |q\xi - p| < 1/2q.$$

Hence

$$\begin{aligned} 1/qq_{n-1} &\leq |qp_{n-1} - pq_{n-1}|/qq_{n-1} \\ &= |p_{n-1}/q_{n-1} - p/q| \\ &\leq |p_{n-1}/q_{n-1} - \xi| + |\xi - p/q| \\ &< 1/2qq_{n-1} + 1/2q^2. \end{aligned}$$

But this implies  $q < q_{n-1}$ , which is a contradiction.  $\square$

As an application of Proposition 4 we prove

**Proposition 5** *Let  $d$  be a positive integer which is not a square and  $m$  an integer such that  $0 < m^2 < d$ . If  $x, y$  are positive integers such that*

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

*then  $x/y$  is a convergent of the irrational number  $\sqrt{d}$ .*

*Proof* Suppose first that  $m > 0$ . Then  $x/y > \sqrt{d}$  and

$$0 < x/y - \sqrt{d} = m/(xy + y^2\sqrt{d}) < \sqrt{d}/2y^2\sqrt{d} = 1/2y^2.$$

Hence  $x/y$  is a convergent of  $\sqrt{d}$ , by Proposition 4.

Suppose next that  $m < 0$ . Then  $y/x > 1/\sqrt{d}$  and

$$0 < y/x - 1/\sqrt{d} = -m/d(xy + x^2/\sqrt{d}) < 1/\sqrt{d}(xy + x^2/\sqrt{d}) < 1/2x^2.$$

Hence  $y/x$  is a convergent of  $1/\sqrt{d}$ . But, since  $1/\sqrt{d} = 0 + 1/\sqrt{d}$ , the convergents of  $1/\sqrt{d}$  are  $0/1$  and the reciprocals of the convergents of  $\sqrt{d}$ .  $\square$

In the next section we will show that the continued fraction expansion of  $\sqrt{d}$  has a particularly simple form.

It was shown by Vahlen (1895) that at least one of any two consecutive convergents of  $\xi$  satisfies the inequality of Proposition 4. Indeed, since consecutive convergents lie on opposite sides of  $\xi$ ,

$$\begin{aligned} |p_n/q_n - \xi| + |p_{n-1}/q_{n-1} - \xi| &= |p_n/q_n - p_{n-1}/q_{n-1}| \\ &= 1/q_n q_{n-1} \leq 1/2q_n^2 + 1/2q_{n-1}^2, \end{aligned}$$

with equality only if  $q_n = q_{n-1}$ . This proves the assertion, except when  $n = 1$  and  $q_1 = q_0 = 1$ . But in this case  $a_1 = 1$ ,  $1 \leq \xi_1 < 2$  and hence

$$|\xi - p_1/q_1| = |\xi - a_0 - 1| = 1 - \xi_1^{-1} < 1/2.$$

It was shown by Borel (1903) that at least one of any three consecutive convergents of  $\xi$  satisfies the sharper inequality

$$|\xi - p/q| < 1/\sqrt{5}q^2.$$

In fact this is obtained by taking  $r = 1$  in the following more general result, due to Forder (1963) and Wright (1964).

**Proposition 6** *Let  $\xi$  be an irrational number with the continued fraction expansion  $[a_0, a_1, \dots]$  and the convergents  $p_n/q_n$ . If, for some positive integer  $r$ ,*

$$|\xi - p_n/q_n| \geq 1/(r^2 + 4)^{1/2} q_n^2 \quad \text{for } n = m - 1, m, m + 1,$$

*then  $a_{m+1} < r$ .*

*Proof* If we put  $s = (r^2 + 4)^{1/2}/2$ , then  $s$  is irrational. For otherwise  $2s$  would be an integer and from  $(2s + r)(2s - r) = 4$  we would obtain  $2s + r = 4$ ,  $2s - r = 1$  and hence  $r = 3/2$ , which is a contradiction.

By the hypotheses of the proposition,

$$\begin{aligned} 1/q_{m-1}q_m &= |p_{m-1}/q_{m-1} - p_m/q_m| = |\xi - p_{m-1}/q_{m-1}| + |\xi - p_m/q_m| \\ &\geq (q_{m-1}^{-2} + q_m^{-2})/2s \end{aligned}$$

and hence

$$q_m^2 - 2sq_{m-1}q_m + q_{m-1}^2 \leq 0.$$

Furthermore, this inequality also holds when  $q_{m-1}, q_m$  are replaced by  $q_m, q_{m+1}$ . Consequently  $q_{m-1}/q_m$  and  $q_{m+1}/q_m$  both satisfy the inequality  $t^2 - 2st + 1 \leq 0$ . Since

$$t^2 - 2st + 1 = (t - s + r/2)(t - s - r/2),$$

it follows that

$$s - r/2 < q_{m-1}/q_m < q_{m+1}/q_m < s + r/2,$$

the first and last inequalities being strict because  $s$  is irrational. Hence

$$a_{m+1} = q_{m+1}/q_m - q_{m-1}/q_m < s + r/2 - (s - r/2) = r. \quad \square$$

It follows from Proposition 6 with  $r = 1$  that, for any irrational number  $\xi$ , there exist infinitely many rational numbers  $p/q = p_n/q_n$  such that

$$|\xi - p/q| < 1/\sqrt{5}q^2.$$

Here the constant  $\sqrt{5}$  is best possible. For take any  $c > \sqrt{5}$ . If there exists a rational number  $p/q$ , with  $q > 0$  and  $(p, q) = 1$ , such that

$$|\xi - p/q| < 1/cq^2,$$

then  $p/q$  is a convergent of  $\xi$ , by Proposition 4. But for any convergent  $p_n/q_n$  we have

$$|\xi - p_n/q_n| = 1/q_n(q_n\xi_{n+1} + q_{n-1}).$$

If we take  $\xi = \tau := (1 + \sqrt{5})/2$ , then also  $\xi_{n+1} = \tau$  and  $p_n = q_{n+1}$ . Hence

$$|\tau - q_{n+1}/q_n| = 1/q_n^2(\tau + q_{n-1}/q_n),$$

where  $\tau + q_{n-1}/q_n \rightarrow \tau + \tau^{-1} = \sqrt{5}$ , since  $q_n/q_{n-1} \rightarrow \tau$ . Thus, for any  $c > \sqrt{5}$ , there exist at most finitely many rational numbers  $p/q$  such that

$$|\tau - p/q| < 1/cq^2.$$

It follows from Proposition 6 with  $r = 2$  that if

$$|\xi - p_n/q_n| \geq 1/\sqrt{8}q_n^2 \quad \text{for all large } n,$$

then  $a_n = 1$  for all large  $n$ . The constant  $\sqrt{8}$  is again best possible, since a similar argument to that just given shows that if  $\sigma := 1 + \sqrt{2} = [2, 2, \dots]$  then, for any  $c > \sqrt{8}$ , there exist at most finitely many rational numbers  $p/q$  such that

$$|\sigma - p/q| < 1/cq^2.$$

It follows from Proposition 6 with  $r = 3$  that if

$$|\xi - p_n/q_n| \geq 1/\sqrt{13}q_n^2 \quad \text{for all large } n,$$

then  $a_n \in \{1, 2\}$  for all large  $n$ .

For any irrational  $\xi$ , with continued fraction expansion  $[a_0, a_1, \dots]$  and convergents  $p_n/q_n$ , put

$$M(\xi) = \overline{\lim}_{n \rightarrow \infty} q_n^{-1} |q_n \xi - p_n|^{-1}.$$

It follows from Proposition 2 that  $M(\xi) = M(\eta)$  if  $\xi$  and  $\eta$  are equivalent. The results just established show that  $M(\xi) \geq \sqrt{5}$  for every  $\xi$ . If  $M(\xi) < \sqrt{8}$ , then  $a_n = 1$  for all large  $n$ ; hence  $\xi$  is equivalent to  $\tau$  and  $M(\xi) = M(\tau) = \sqrt{5}$ . If  $M(\xi) < \sqrt{13}$ , then  $a_n \in \{1, 2\}$  for all large  $n$ .

An irrational number  $\xi$  is said to be *badly approximable* if  $M(\xi) < \infty$ . The inequalities

$$a_{n+2}/q_n q_{n+2} < |\xi - p_n/q_n| < 1/q_n q_{n+1}$$

imply

$$a_{n+1} \leq q_{n+1}/q_n < q_n^{-1} |q_n \xi - p_n|^{-1}$$

and

$$q_n^{-1} |q_n \xi - p_n|^{-1} < q_{n+2}/a_{n+2} q_n \leq q_{n+1}/q_n + 1 \leq a_{n+1} + 2.$$

Hence  $\xi$  is badly approximable if and only if its partial quotients  $a_n$  are bounded.

It is obvious that  $\xi$  is badly approximable if there exists a constant  $c > 0$  such that

$$|\xi - p/q| > c/q^2$$

for every rational number  $p/q$ . Conversely, if  $\xi$  is badly approximable, then there exists such a constant  $c > 0$ . This is clear when  $p$  and  $q$  are coprime integers, since if  $p/q$  is not a convergent of  $\xi$  then, by Proposition 4,

$$|\xi - p/q| \geq 1/2q^2.$$

On the other hand, if  $p = \lambda p'$ ,  $q = \lambda q'$ , where  $p', q'$  are coprime, then

$$|\xi - p/q| = |\xi - p'/q'| \geq c/q'^2 = \lambda^2 c/q^2 \geq c/q^2.$$

Some of the applications of badly approximable numbers stem from the following characterization: a real number  $\theta$  is badly approximable if and only if there exists a constant  $c' > 0$  such that

$$|e^{2\pi i q \theta} - 1| \geq c'/q \quad \text{for all } q \in \mathbb{N}.$$

To establish this, put  $q\theta = p + \delta$ , where  $p \in \mathbb{Z}$  and  $|\delta| \leq 1/2$ . Then

$$|e^{2\pi i q\theta} - 1| = 2|\sin \pi q\theta| = 2|\sin \pi \delta|$$

and the result follows from the previous characterization, since  $(\sin x)/x$  decreases from 1 to  $2/\pi$  as  $x$  increases from 0 to  $\pi/2$ .

### 3 Periodic Continued Fractions

A complex number  $\zeta$  is said to be a *quadratic irrational* if it is a root of a monic quadratic polynomial  $t^2 + rt + s$  with rational coefficients  $r, s$ , but is not itself rational. Since  $\zeta \notin \mathbb{Q}$ , the rational numbers  $r, s$  are uniquely determined by  $\zeta$ .

Equivalently,  $\zeta$  is a quadratic irrational if it is a root of a quadratic polynomial

$$f(t) = At^2 + Bt + C$$

with integer coefficients  $A, B, C$  such that  $B^2 - 4AC$  is not the square of an integer. The integers  $A, B, C$  are uniquely determined up to a common factor and are uniquely determined up to sign if we require that they have greatest common divisor 1. The corresponding integer  $D = B^2 - 4AC$  is then uniquely determined and is called the *discriminant* of  $\zeta$ . A quadratic irrational is real if and only if its discriminant is positive.

It is readily verified that if a quadratic irrational  $\zeta$  is equivalent to a complex number  $\omega$ , i.e. if

$$\zeta = (\alpha\omega + \beta)/(\gamma\omega + \delta),$$

where  $\alpha, \beta, \gamma, \delta \in \mathbb{Z}$  and  $\alpha\delta - \beta\gamma = \pm 1$ , then  $\omega$  is also a quadratic irrational. Moreover, if  $\zeta$  is a root of the quadratic polynomial  $f(t) = At^2 + Bt + C$ , where  $A, B, C$  are integers with greatest common divisor 1, then  $\omega$  is a root of the quadratic polynomial

$$g(t) = A't^2 + B't + C',$$

where

$$\begin{aligned} A' &= \alpha^2 A + \alpha\gamma B + \gamma^2 C, \\ B' &= 2\alpha\beta A + (\alpha\delta + \beta\gamma)B + 2\gamma\delta C, \\ C' &= \beta^2 A + \beta\delta B + \delta^2 C, \end{aligned}$$

and hence

$$B'^2 - 4A'C' = B^2 - 4AC = D.$$

Since

$$\begin{aligned} A &= \delta^2 A' - \gamma\delta B' + \gamma^2 C', \\ B &= -2\beta\delta A' + (\alpha\delta + \beta\gamma)B' - 2\alpha\gamma C', \\ C &= \beta^2 A' - \alpha\beta B' + \alpha^2 C', \end{aligned}$$

$A', B', C'$  also have greatest common divisor 1.

If  $\zeta$  is a quadratic irrational, we define the *conjugate*  $\zeta'$  of  $\zeta$  to be the other root of the quadratic polynomial  $f(t)$  which has  $\zeta$  as a root. If

$$\zeta = (\alpha\omega + \beta)/(\gamma\omega + \delta),$$

where  $\alpha, \beta, \gamma, \delta \in \mathbb{Z}$  and  $\alpha\delta - \beta\gamma = \pm 1$ , then evidently

$$\zeta' = (\alpha\omega' + \beta)/(\gamma\omega' + \delta).$$

Suppose now that  $\zeta = \zeta'$  is real and that the integers  $A, B, C$  are uniquely determined by requiring not only  $(A, B, C) = 1$  but also  $A > 0$ . The real quadratic irrational  $\zeta$  is said to be *reduced* if  $\zeta > 1$  and  $-1 < \zeta' < 0$ . If  $\zeta$  is reduced then, since  $\zeta > \zeta'$ , we must have

$$\zeta = (-B + \sqrt{D})/2A, \quad \zeta' = (-B - \sqrt{D})/2A.$$

Thus the inequalities  $\zeta > 1$  and  $-1 < \zeta' < 0$  imply

$$0 < \sqrt{D} + B < 2A < \sqrt{D} - B.$$

Conversely, if the coefficients  $A, B, C$  of  $f(t)$  satisfy these inequalities, where  $D = B^2 - 4AC > 0$ , then one of the roots of  $f(t)$  is reduced. For  $B < 0 < A$  and so the roots  $\zeta, \zeta'$  of  $f(t)$  have opposite signs. If  $\zeta$  is the positive root, then  $\zeta$  and  $\zeta'$  are given by the preceding formulas and hence  $\zeta > 1, -1 < \zeta' < 0$ . It should be noted also that if  $\zeta$  is reduced, then  $B^2 < D$  and hence  $C < 0$ .

We return now to continued fractions. If  $\zeta$  is a real quadratic irrational, then its complete quotients  $\zeta_n$  are all quadratic irrationals and, conversely, if some complete quotient  $\zeta_n$  is a quadratic irrational, then  $\zeta$  is also a quadratic irrational.

The continued fraction expansion  $[a_0, a_1, a_2, \dots]$  of a real number  $\zeta$  is said to be *eventually periodic* if there exist integers  $m \geq 0$  and  $h > 0$  such that

$$a_n = a_{n+h} \quad \text{for all } n \geq m.$$

The continued fraction expansion is then conveniently denoted by

$$[a_0, a_1, \dots, a_{m-1}, \overline{a_m, \dots, a_{m+h-1}}].$$

The continued fraction expansion is said to be *periodic* if it is eventually periodic with  $m = 0$ .

Equivalently, the continued fraction expansion of  $\zeta$  is eventually periodic if  $\zeta_m = \zeta_{m+h}$  for some  $m \geq 0$  and  $h > 0$ , and periodic if this holds with  $m = 0$ . The *period* of the continued fraction expansion, in either case, is the least positive integer  $h$  with this property.

We are going to show that there is a close connection between real quadratic irrationals and eventually periodic continued fractions.

**Proposition 7** *A real number  $\zeta$  is a reduced quadratic irrational if and only if its continued fraction expansion is periodic.*

*Moreover, if  $\zeta = [a_0, \dots, a_{h-1}]$ , then  $-1/\zeta' = [\overline{a_{h-1}, \dots, a_0}]$ .*

*Proof* Suppose first that  $\xi = [\overline{a_0, \dots, a_{h-1}}]$  has a periodic continued fraction expansion. Then  $a_0 = a_h \geq 1$  and hence  $\xi > 1$ . Furthermore, since

$$\xi = (p_{h-1}\xi_h + p_{h-2})/(q_{h-1}\xi_h + q_{h-2})$$

and  $\xi_h = \xi$ ,  $\xi$  is an irrational root of the quadratic polynomial

$$f(t) = q_{h-1}t^2 + (q_{h-2} - p_{h-1})t - p_{h-2}.$$

Thus  $\xi$  is a quadratic irrational. Since  $f(0) = -p_{h-2} < 0$  and

$$f(-1) = q_{h-1} - q_{h-2} + p_{h-1} - p_{h-2} > 0$$

(even for  $h = 1$ ), it follows that  $-1 < \xi' < 0$ . Thus  $\xi$  is reduced.

If  $\xi$  is a reduced quadratic irrational, then its complete quotients  $\xi_n$ , which are all quadratic irrationals, are also reduced, by Lemma 0 with  $\eta = \xi'$ . Since  $\xi'_n = a_n + 1/\xi'_{n+1}$  and  $-1 < \xi'_n < 0$ , we have

$$a_n = \lfloor -1/\xi'_{n+1} \rfloor.$$

Thus  $\xi_n, \xi'_n$  are the roots of a uniquely determined polynomial

$$f_n(t) = A_nt^2 + B_nt + C_n,$$

where  $A_n, B_n, C_n$  are integers with greatest common divisor 1 and  $A_n > 0$ . Furthermore,  $D = B_n^2 - 4A_nC_n$  is independent of  $n$  and positive. Since  $\xi_n$  is reduced, we have

$$\xi_n = (-B_n + \sqrt{D})/2A_n, \quad \xi'_n = (-B_n - \sqrt{D})/2A_n,$$

where

$$0 < \sqrt{D} + B_n < 2A_n < \sqrt{D} - B_n.$$

If we put  $g = \lfloor \sqrt{D} \rfloor$ , then  $-B_n \in \{1, \dots, g\}$  and, for a given value of  $B_n$ , there are at most  $-B_n$  possible values for  $A_n$ . Consequently the number of distinct pairs  $A_n, B_n$  does not exceed  $1 + \dots + g = g(g+1)/2$ . Hence we must have

$$\xi_j = \xi_k, \quad \xi'_j = \xi'_k$$

for some  $j, k$  such that  $0 \leq j < k \leq g(g+1)/2$ . If  $j = 0$ , this already proves that the continued fraction expansion of  $\xi$  is periodic. If  $j > 0$ , then

$$a_{j-1} = \lfloor -1/\xi'_j \rfloor = \lfloor -1/\xi'_k \rfloor = a_{k-1}$$

and hence

$$\xi_{j-1} = a_{j-1} + 1/\xi_j = a_{k-1} + 1/\xi_k = \xi_{k-1}.$$

Repeating this argument  $j$  times, we obtain  $\xi_0 = \xi_{k-j}$ . Thus  $\xi$  has a periodic continued fraction expansion in any case.

If the period is  $h$ , so that  $\xi = [\overline{a_0, \dots, a_{h-1}}]$ , then  $\xi'_0 = \xi'_h$  and the relation  $a_n = \lfloor -1/\xi'_{n+1} \rfloor$  implies that  $-1/\xi' = [\overline{a_{h-1}, \dots, a_0}]$ .  $\square$

The proof of Proposition 7 shows that the period is at most  $g(g+1)/2$  and thus is certainly less than  $D$ . By counting the pairs of integers  $A, B$  for which not only

$$0 < \sqrt{D} + B < 2A < \sqrt{D} - B,$$

but also  $D \equiv B^2 \pmod{4A}$ , it may be shown that the period is at most  $O(\sqrt{D} \log D)$ . (The *Landau order symbol* used here is defined under 'Notations'.)

**Proposition 8** *A real number  $\xi$  is a quadratic irrational if and only if its continued fraction expansion is eventually periodic.*

*Proof* Suppose first that the continued fraction expansion of  $\xi$  is eventually periodic. Then some complete quotient  $\xi_m$  has a periodic continued fraction expansion and hence is a quadratic irrational, by Proposition 7. But this implies that  $\xi$  also is a quadratic irrational.

Suppose next that  $\xi$  is a quadratic irrational. We will prove that the continued fraction expansion of  $\xi$  is eventually periodic by showing that some complete quotient  $\xi_{n+1}$  is reduced. Since we certainly have  $\xi_{n+1} > 1$ , we need only show that  $-1 < \xi'_{n+1} < 0$ . But  $\xi' \neq \xi$  and  $\xi' = (p_n \xi'_{n+1} + p_{n-1}) / (q_n \xi'_{n+1} + q_{n-1})$ . Hence, by Lemma 0,  $-1 < \xi'_{n+1} < 0$  for all large  $n$ .  $\square$

It follows from Proposition 8 that any real quadratic irrational is badly approximable, since its partial quotients are bounded. It follows from Propositions 7 and 8 that there are only finitely many inequivalent quadratic irrationals with a given discriminant  $D > 0$ , since any real quadratic irrational is equivalent to a reduced one and only finitely many pairs of integers  $A, B$  satisfy the inequalities

$$0 < \sqrt{D} + B < 2A < \sqrt{D} - B.$$

Proposition 8 is due to Euler and Lagrange. It was first shown by Euler (1737) that a real number is a quadratic irrational if its continued fraction expansion is eventually periodic, and the converse was proved by Lagrange (1770). Proposition 7 was first stated and proved by Galois (1829), although it was implicit in the work of Lagrange (1773) on the reduction of binary quadratic forms. Proposition 7 provides a simple proof of the following result due to Legendre:

**Proposition 9** *For any real number  $\xi$ , the following two conditions are equivalent:*

- (i)  $\xi > 1$ ,  $\xi$  is irrational and  $\xi^2$  is rational;
- (ii) *the continued fraction expansion of  $\xi$  has the form  $[a_0, \overline{a_1, \dots, a_h}]$ , where  $a_h = 2a_0$  and  $a_i = a_{h-i}$  for  $i = 1, \dots, h-1$ .*

*Proof* Suppose first that (i) holds. Then  $\xi$  is a quadratic irrational, since it is a root of the polynomial  $t^2 - \xi^2$ . The continued fraction expansion of  $\xi$  cannot be periodic, by Proposition 7, since  $\xi' = -\xi < -1$ . However, the continued fraction expansion of  $\xi_1$  is periodic, since  $\xi_1 > 1$  and  $1/\xi'_1 = \xi' - a_0 < -1$ . Thus  $\xi_1 = [\overline{a_1, \dots, a_h}]$  for some  $h \geq 1$ . By Proposition 7 also,

$$-1/\xi'_1 = [\overline{a_h, \dots, a_1}].$$

But

$$-1/\zeta'_1 = \zeta + a_0 = [2a_0, \overline{a_1, \dots, a_h}].$$

Comparing this with the previous expression, we see that (ii) holds.

Suppose, conversely, that (ii) holds. Then  $\zeta$  is irrational,  $a_0 > 0$  and hence  $\zeta > 1$ . Moreover  $\zeta_1 = [\overline{a_1, \dots, a_h}]$  is a reduced quadratic irrational and

$$-1/\zeta'_1 = [\overline{a_h, \dots, a_1}] = [2a_0, \overline{a_1, \dots, a_h}] = a_0 + \zeta.$$

Hence  $\zeta' = a_0 + 1/\zeta'_1 = -\zeta$  and  $\zeta^2 = -\zeta\zeta'$  is rational.  $\square$

## 4 Quadratic Diophantine Equations

We are interested in finding all integers  $x, y$  such that

$$ax^2 + bxy + cy^2 + dx + ey + f = 0, \quad (6)$$

where  $a, \dots, f$  are given integers. Writing (6) as a quadratic equation for  $x$ ,

$$ax^2 + (by + d)x + cy^2 + ey + f = 0,$$

we see that if a solution exists for some  $y$ , then the discriminant

$$(by + d)^2 - 4a(cy^2 + ey + f)$$

must be a perfect square. Thus

$$(b^2 - 4ac)y^2 + 2(bd - 2ae)y + d^2 - 4af = z^2$$

for some integer  $z$ . If we put

$$p := b^2 - 4ac, \quad q := bd - 2ae, \quad r := d^2 - 4af,$$

we have a quadratic equation for  $y$ ,

$$py^2 + 2qy + r - z^2 = 0,$$

whose discriminant must also be a perfect square. Thus

$$q^2 - p(r - z^2) = w^2$$

for some integer  $w$ . Thus if (6) has a solution in integers, so also does the equation

$$w^2 - pz^2 = q^2 - pr.$$

Moreover, from all solutions in integers of the latter equation we may obtain, by retracing our steps, all solutions in integers of the original equation (6).

Thus we now restrict our attention to finding all integers  $x, y$  such that

$$x^2 - dy^2 = m, \quad (7)$$

where  $d$  and  $m$  are given integers.

The equation (7) has the remarkable property, which was known to Brahmagupta (628) and later rediscovered by Euler (1758), that if we have solutions for two values  $m_1, m_2$  of  $m$ , then we can derive a solution for their product  $m_1 m_2$ . This follows from the identity

$$(x_1^2 - dy_1^2)(x_2^2 - dy_2^2) = x^2 - dy^2,$$

where

$$x = x_1x_2 + dy_1y_2, \quad y = x_1y_2 + y_1x_2.$$

(In fact, Brahmagupta's identity is just a restatement of the norm relation  $N(\alpha\beta) = N(\alpha)N(\beta)$  for elements  $\alpha, \beta$  of a quadratic field.) In particular, from two solutions of the equation

$$x^2 - dy^2 = 1, \tag{8}$$

a third solution can be obtained by *composition* in this way.

Composition of solutions is evidently commutative and associative. In fact the solutions of (8) form an abelian group under composition, with the trivial solution 1, 0 as identity element and the solution  $x, -y$  as the inverse of the solution  $x, y$ . Also, by composing an arbitrary solution  $x, y$  of (8) with the trivial solution  $-1, 0$  we obtain the solution  $-x, -y$ .

Suppose first that  $d < 0$ . Evidently (7) is insoluble if  $m < 0$  and  $x = y = 0$  is the only solution if  $m = 0$ . If  $m > 0$ , there are at most finitely many solutions and we may find them all by testing, for each non-negative integer  $y \leq (-m/d)^{1/2}$ , whether  $m + dy^2$  is a perfect square.

Suppose now that  $d > 0$ . If  $d = e^2$  is a perfect square, then (7) is equivalent to the finite set of simultaneous linear Diophantine equations

$$x - ey = m', \quad x + ey = m'',$$

where  $m', m''$  are any integers such that  $m'm'' = m$ . Thus we now suppose also that  $d$  is not a perfect square. Then  $\xi = \sqrt{d}$  is irrational.

If  $0 < m^2 < d$  then, by Proposition 5, any positive solution  $x, y$  of (7) has the form  $x = p_n, y = q_n$ , where  $p_n/q_n$  is a convergent of  $\xi$ . In particular, all positive solutions of  $x^2 - dy^2 = \pm 1$  are obtained in this way.

On the other hand, as we now show, if  $p_n/q_n$  is any convergent of  $\xi$  then

$$|p_n^2 - dq_n^2| < 2\sqrt{d}.$$

If  $n = 0$ , then  $|p_0^2 - dq_0^2| = |a_0^2 - d|$ , where  $a_0 < \sqrt{d} < a_0 + 1$  and so  $0 < d - a_0^2 < \sqrt{d} + a_0 < 2\sqrt{d}$ . Now suppose  $n > 0$ . Then  $|p_n - q_n\xi| < q_{n+1}^{-1}$  and hence

$$\begin{aligned} |p_n^2 - dq_n^2| &= |p_n - q_n\xi||p_n + q_n\xi + 2q_n\xi| \\ &< q_{n+1}^{-1}(q_{n+1}^{-1} + 2q_n\xi) < 2\xi. \end{aligned}$$

An easy congruence argument shows that the equation

$$x^2 - dy^2 = -1 \tag{9}$$

has no solutions in integers unless  $d \equiv 1 \pmod{4}$  or  $d \equiv 2 \pmod{8}$ . It will now be shown that the equation (8), on the other hand, always has solutions in positive integers.

**Proposition 10** Let  $d$  be a positive integer which is not a perfect square. Suppose  $\xi = \sqrt{d}$  has complete quotients  $\xi_n$ , convergents  $p_n/q_n$ , and continued fraction expansion  $[a_0, \overline{a_1, \dots, a_h}]$  of period  $h$ .

Then  $p_n^2 - dq_n^2 = \pm 1$  if and only if  $n = kh - 1$  for some integer  $k > 0$  and in this case

$$p_{kh-1}^2 - dq_{kh-1}^2 = (-1)^{kh}.$$

*Proof* From  $\xi = (p_n \xi_{n+1} + p_{n-1}) / (q_n \xi_{n+1} + q_{n-1})$  we obtain

$$(p_n - q_n \xi) \xi_{n+1} = q_{n-1} \xi - p_{n-1}.$$

Multiplying by  $(-1)^{n+1}(p_n + q_n \xi)$ , we get

$$s_n \xi_{n+1} = \xi + r_n,$$

where

$$s_n = (-1)^{n+1}(p_n^2 - dq_n^2), \quad r_n = (-1)^n(p_{n-1}p_n - dq_{n-1}q_n).$$

Thus  $s_n$  and  $r_n$  are integers. Moreover, since  $\xi_{n+1+kh} = \xi_{n+1}$  and  $\xi$  is irrational,  $s_{n+kh} = s_n$  and  $r_{n+kh} = r_n$  for all positive integers  $k$ .

If  $p_n^2 - dq_n^2 = \pm 1$ , then actually  $p_n^2 - dq_n^2 = (-1)^{n+1}$ , since  $p_n/q_n$  is less than or greater than  $\xi$  according as  $n$  is even or odd. Hence  $s_n = 1$  and  $\xi_{n+1} = \xi + r_n$ . Taking integral parts, we get  $a_{n+1} = a_0 + r_n$ . Consequently

$$\xi_{n+2}^{-1} = \xi_{n+1} - a_{n+1} = \xi - a_0 = \xi_1^{-1}.$$

Thus  $\xi_{n+2} = \xi_1$ , which implies that  $n = kh - 1$  for some positive integer  $k$ .

On the other hand, if  $n = kh - 1$  for some positive integer  $k$ , then  $\xi_{n+2} = \xi_1$  and hence

$$\xi_{n+1} - a_{n+1} = \xi - a_0.$$

Thus  $\xi_{n+1} = \xi + a_{n+1} - a_0$ , which implies that  $s_n = 1$ , since  $\xi$  is irrational.  $\square$

It follows from Proposition 10 that, if  $d$  is a positive integer which is not a perfect square, then the equation (8) always has a solution in positive integers and all such solutions are given by

$$\begin{aligned} x &= p_{kh-1}, & y &= q_{kh-1} & (k = 1, 2, \dots) & \text{if } h \text{ is even,} \\ x &= p_{2kh-1}, & y &= q_{2kh-1} & (k = 1, 2, \dots) & \text{if } h \text{ is odd.} \end{aligned}$$

The least solution in positive integers, obtained by taking  $k = 1$ , is called the *fundamental solution* of (8).

On the other hand, the equation (9) has a solution in positive integers if and only if  $h$  is odd and all solutions are then given by

$$x = p_{kh-1}, \quad y = q_{kh-1} \quad (k = 1, 3, 5, \dots).$$

The least solution in positive integers, obtained by taking  $k = 1$ , is called the *fundamental solution* of (9).

To illustrate these results, suppose  $d = a^2 + 1$  for some  $a \in \mathbb{N}$ . Since  $\sqrt{d} = [a, \overline{2a}]$ , the equation  $x^2 - dy^2 = -1$  has the fundamental solution  $x = a, y = 1$  and the equation  $x^2 - dy^2 = 1$  has the fundamental solution  $x = 2a^2 + 1, y = 2a$ . Again, suppose  $d = a^2 + a$  for some  $a \in \mathbb{N}$ . Since  $\sqrt{d} = [a, \overline{2, 2a}]$ , the equation  $x^2 - dy^2 = -1$  is insoluble, but the equation  $x^2 - dy^2 = 1$  has the fundamental solution  $x = 2a + 1, y = 2$ .

It is not difficult to obtain upper bounds for the fundamental solutions. Since  $\xi = \sqrt{d}$  is a root of the polynomial  $t^2 - d$  and since its complete quotients  $\xi_n$  are reduced for  $n \geq 1$ , they have the form

$$\xi_n = (-B_n + \sqrt{D})/2A_n,$$

where  $D = 4d, 0 < -B_n < \sqrt{D}$  and  $A_n \geq 1$ . Therefore  $a_0 = \lfloor \xi \rfloor < \sqrt{d}$  and  $a_n = \lfloor \xi_n \rfloor < 2\sqrt{d}$  for  $n \geq 1$ . If we put  $\alpha = \lfloor \sqrt{d} \rfloor$ , it is easily shown by induction that

$$p_n \leq (\alpha + \alpha^{-1})^{n+1}/2, \quad q_n \leq (\alpha + \alpha^{-1})^n \quad (n \geq 0).$$

These inequalities may now be combined with any upper bound for the period  $h$  (cf. §3).

Under composition, the fundamental solution of (8) generates an infinite cyclic group  $\mathcal{C}$  of solutions of (8). Furthermore, by composing the fundamental solution of (9) with any element of  $\mathcal{C}$  we obtain infinitely many solutions of (9). We are going to show that, by composing also with the trivial solution  $-1, 0$  of (8), all integral solutions of (8) and (9) are obtained in this way. This can be proved by means of continued fractions, but the following argument due to Nagell (1950) provides additional information.

**Proposition 11** *Let  $d$  be a positive integer which is not a perfect square, let  $m$  be a positive integer, and let  $x_0, y_0$  be the fundamental solution of the equation (8).*

*If the equation*

$$u^2 - dv^2 = m \tag{10}$$

*has an integral solution, then it actually has one for which  $u^2 \leq m(x_0 + 1)/2$ ,  $dv^2 \leq m(x_0 - 1)/2$ .*

*Similarly, if the equation*

$$u^2 - dv^2 = -m \tag{11}$$

*has an integral solution, then it actually has one for which  $u^2 \leq m(x_0 - 1)/2$ ,  $dv^2 \leq m(x_0 + 1)/2$ .*

*Proof* By composing a given solution of (10) with any solution in the subgroup  $\mathcal{C}$  of solutions of (8) which is generated by the solution  $x_0, y_0$  we obtain again a solution of (10). Let  $u_0, v_0$  be the solution of (10) obtained in this way for which  $v_0$  has its least non-negative value. Then  $u_0^2 = m + dv_0^2$  also has its least value and by changing the sign of  $u_0$  we may suppose  $u_0 > 0$ . By composing the solution  $u_0, v_0$  of (10) with the inverse of the fundamental solution  $x_0, y_0$  of (8) we obtain the solution

$$u = x_0 u_0 - d y_0 v_0, \quad v = x_0 v_0 - y_0 u_0$$

of (10). Since

$$u = x_0 u_0 - d y_0 v_0 = x_0 u_0 - [(x_0^2 - 1)(u_0^2 - m)]^{1/2} > 0,$$

we must have

$$x_0 u_0 - d y_0 v_0 \geq u_0.$$

Hence

$$(x_0 - 1)^2 u_0^2 \geq d^2 y_0^2 v_0^2 = (x_0^2 - 1)(u_0^2 - m).$$

Thus

$$(x_0 - 1)/(x_0 + 1) \geq 1 - m/u_0^2,$$

which implies  $u_0^2 \leq m(x_0 + 1)/2$  and hence  $d v_0^2 \leq m(x_0 - 1)/2$ .

For the equation (11) we begin in the same way. Then from

$$(x_0 v_0)^2 = (y_0^2 + 1/d)(u_0^2 + m) > y_0^2 u_0^2$$

we obtain  $v = x_0 v_0 - y_0 u_0 > 0$  and hence  $x_0 v_0 - y_0 u_0 \geq v_0$ . Thus

$$d(x_0 - 1)^2 v_0^2 \geq d y_0^2 u_0^2$$

and hence

$$(x_0 - 1)^2(u_0^2 + m) \geq (x_0^2 - 1)u_0^2.$$

The argument can now be completed in the same way as before.  $\square$

The proof of Proposition 11 shows that if (10), or (11), has an integral solution, then we obtain all solutions by finding the *finitely many* solutions  $u, v$  which satisfy the inequalities in the statement of Proposition 11 and composing them with all solutions in  $\mathcal{C}$  of (8).

The only solutions  $x, y$  of (8) for which  $x^2 \leq (x_0 + 1)/2$  are the trivial ones  $x = \pm 1, y = 0$ . Hence any solution of (8) is in  $\mathcal{C}$  or is obtained by reversing the signs of a solution in  $\mathcal{C}$ .

If  $u, v$  is a positive solution of (9) such that  $u^2 \leq (x_0 - 1)/2, d v^2 \leq (x_0 + 1)/2$ , then  $x = u^2 + d v^2, y = 2uv$  is a positive solution of (8) such that  $x \leq x_0$ . Hence  $(x, y) = (x_0, y_0)$  is the fundamental solution of (8) and  $u^2 = (x_0 - 1)/2, d v^2 = (x_0 + 1)/2$ . Thus  $(u, v)$  is uniquely determined and is the fundamental solution of (9). Hence, if (9) has a solution, any solution is obtained by composing the fundamental solution of (9) with an element of  $\mathcal{C}$  or by reversing the signs of such a solution.

A necessary condition for the solubility in integers of the equation (9) is that  $d$  may be represented as a sum of two squares. For the period  $h$  of the continued fraction expansion  $\zeta = \sqrt{d} = [a_0, \overline{a_1, \dots, a_h}]$  must be odd, say  $h = 2m + 1$ . It follows from Proposition 9 that

$$\zeta_{m+1} = [\overline{a_m, \dots, a_1, 2a_0, a_1, \dots, a_m}],$$

and then from Proposition 7 that  $\xi_{m+1} = -1/\xi'_{m+1}$ . But, by the proof of Proposition 10,

$$s_m \xi_{m+1} = \zeta + r_m,$$

where  $s_m$  and  $r_m$  are integers. Hence

$$-1 = \xi_{m+1} \xi'_{m+1} = (\zeta + r_m)(-\zeta + r_m)/s_m^2 = (r_m^2 - d)/s_m^2,$$

and thus  $d = r_m^2 + s_m^2$ . The formulas for  $s_m$  and  $r_m$  show that, if  $p_n/q_n$  are the convergents of  $\sqrt{d}$ , then  $d = x^2 + y^2$  with

$$x = p_{m-1}p_m - dq_{m-1}q_m, \quad y = p_m^2 - dq_m^2.$$

Unfortunately, the equation (9) may be insoluble, even though  $d$  is a sum of two squares. As an example, take  $d = 34 = 5^2 + 3^2$ . It is easily verified that the fundamental solution of the equation  $x^2 - 34y^2 = 1$  is  $x_0 = 35$ ,  $y_0 = 6$ . If the equation  $u^2 - 34v^2 = -1$  were soluble in integers, then, by Proposition 11, it would have a solution  $u, v$  such that  $34v^2 \leq 18$ , which is clearly impossible.

As already observed, the equation (9) has no integral solutions if  $d \equiv 3 \pmod{4}$ . It will now be shown that (9) does have integral solutions if  $d = p$  is prime and  $p \equiv 1 \pmod{4}$ . For let  $x, y$  be the fundamental solution of the equation (8). Since any square is congruent to 0 or 1 mod 4, we must have  $y^2 \equiv 0$  and  $x^2 \equiv 1$ . Thus  $y = 2z$  for some positive integer  $z$  and

$$(x-1)(x+1) = 4pz^2.$$

Since  $x$  is odd,  $x-1$  and  $x+1$  have greatest common divisor 2. It follows that there exist positive integers  $u, v$  such that

$$\text{either } x-1 = 2pu^2, \ x+1 = 2v^2 \quad \text{or} \quad x-1 = 2u^2, \ x+1 = 2pv^2.$$

In the first case  $v^2 - pu^2 = 1$ , which contradicts the choice of  $x, y$  as the fundamental solution of (8), since  $v < x$ . Thus only the second case is possible and then  $u^2 - pv^2 = -1$ . (In fact,  $u, v$  is the fundamental solution of (9).)

This proves again that *any prime  $p \equiv 1 \pmod{4}$  may be represented as a sum of two squares*, and moreover shows that an explicit construction for this representation is provided by the continued fraction expansion of  $\sqrt{p}$ .

The representation of a prime  $p \equiv 1 \pmod{4}$  in the form  $x^2 + y^2$  is actually unique, apart from interchanging  $x$  and  $y$  and changing their signs. For suppose

$$x^2 + y^2 = p = u^2 + v^2,$$

where  $x, y, u, v$  are all positive integers. Then

$$y^2u^2 - x^2v^2 = (p - x^2)u^2 - x^2(p - u^2) = p(u^2 - x^2).$$

Hence  $yu \equiv \varepsilon xv \pmod{p}$ , where  $\varepsilon = \pm 1$ . On the other hand,

$$p^2 = (x^2 + y^2)(u^2 + v^2) = (xu + \varepsilon yv)^2 + (xv - \varepsilon yu)^2.$$

Since the second term on the right is divisible by  $p^2$ , we must have  $xv = \varepsilon yu$  or  $xu = -\varepsilon yv$ . Evidently  $\varepsilon = 1$  in the first case and  $\varepsilon = -1$  in the second case. Since  $(x, y) = (u, v) = 1$ , it follows that either  $x = u, y = v$  or  $x = v, y = u$ .

The equation  $x^2 - dy^2 = 1$ , where  $d$  is a positive integer which is not a perfect square, is generally known as *Pell's equation*, following an erroneous attribution of Euler. The problem of finding its integral solutions was issued as a challenge by Fermat (1657). In the same year Brouncker and Wallis gave a method of solution which is essentially the same as the solution by continued fractions. The first complete proof that a nontrivial solution always exists was given by Lagrange (1768).

Unknown to them all, the problem had been considered centuries earlier by Hindu mathematicians. Special cases of Pell's equation were solved by Brahmagupta (628) and a general method of solution, which was described by Bhascara II (1150), was known to Jayadeva at least a century earlier. No proofs were given, but their method is a modification of the solution by continued fractions and is often faster in practice. Bhascara found the fundamental solution of the equation  $x^2 - 61y^2 = 1$ , namely

$$x = 1766319049, \quad y = 226153980,$$

a remarkable achievement for the era.

## 5 The Modular Group

We recall that a complex number  $w$  is said to be *equivalent* to a complex number  $z$  if there exist integers  $a, b, c, d$  with  $ad - bc = \pm 1$  such that

$$w = (az + b)/(cz + d).$$

Since we can write

$$w = (az + b)(c\bar{z} + d)/|cz + d|^2,$$

the imaginary parts are related by

$$\mathcal{I}w = (ad - bc)\mathcal{I}z/|cz + d|^2.$$

Consequently  $\mathcal{I}w$  and  $\mathcal{I}z$  have the same sign if  $ad - bc = 1$  and opposite signs if  $ad - bc = -1$ . Since the map  $z \rightarrow -z$  interchanges the upper and lower half-planes, we may restrict attention to  $z$ 's in the *upper half-plane*  $\mathcal{H} = \{z \in \mathbb{C} : \mathcal{I}z > 0\}$  and to  $w$ 's which are *properly equivalent* to them, i.e. with  $ad - bc = 1$ .

A *modular transformation* is a map  $f : \mathcal{H} \rightarrow \mathcal{H}$  of the form

$$f(z) = (az + b)/(cz + d),$$

where  $a, b, c, d \in \mathbb{Z}$  and  $ad - bc = 1$ . Such a map is bijective and its inverse is again a modular transformation:

$$f^{-1}(z) = (dz - b)/(-cz + a).$$

Furthermore, if

$$g(z) = (a'z + b')/(c'z + d')$$

is another modular transformation, then the composite map  $h = g \circ f$  is again a modular transformation:

$$h(z) = (a''z + b'')/(c''z + d''),$$

where

$$\begin{aligned} a'' &= a'a + b'c, & b'' &= a'b + b'd, \\ c'' &= c'a + d'c, & d'' &= c'b + d'd, \end{aligned}$$

and hence

$$a''d'' - b''c'' = (a'd' - b'c')(ad - bc) = 1.$$

It follows that the set  $\Gamma$  of all modular transformations is a group. Moreover, composition of modular transformations corresponds to multiplication of the corresponding matrices:

$$\begin{pmatrix} a'' & b'' \\ c'' & d'' \end{pmatrix} = \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

However, the same modular transformation is obtained if the signs of  $a, b, c, d$  are all changed (and in no other way). It follows that the modular group  $\Gamma$  is isomorphic to the factor group  $SL_2(\mathbb{Z})/\{\pm I\}$  of the special linear group  $SL_2(\mathbb{Z})$  of all  $2 \times 2$  integer matrices with determinant 1 by its centre  $\{\pm I\}$ .

**Proposition 12** *The modular group  $\Gamma$  is generated by the transformations*

$$T(z) = z + 1, \quad S(z) = -1/z.$$

*Proof* It is evident that  $S, T \in \Gamma$  and  $S^2 = I$  is the identity transformation. Any  $g \in \Gamma$  has the form

$$g(z) = (az + b)/(cz + d),$$

where  $a, b, c, d \in \mathbb{Z}$  and  $ad - bc = 1$ . If  $c = 0$ , then  $a = d = \pm 1$  and  $g = T^m$ , where  $m = b/d \in \mathbb{Z}$ . Similarly if  $a = 0$ , then  $b = -c = \pm 1$  and  $g = ST^m$ , where  $m = d/c \in \mathbb{Z}$ . Suppose now that  $ac \neq 0$ . For any  $n \in \mathbb{Z}$  we have

$$ST^{-n}g(z) = (a'z + b')/(c'z + d'),$$

where  $a' = -c, b' = -d, c' = a - nc$  and  $d' = b - nd$ . We can choose  $n = m_1$  so that for  $g_1 = ST^{-m_1}g$  we have  $|c'| < |a|$  and hence  $|a'| + |c'| < |a| + |c|$ . If  $a'c' \neq 0$ , the argument can be repeated with  $g_1$  in place of  $g$ . After finitely many repetitions we must obtain

$$ST^{-m_k} \dots ST^{-m_1}g = T^m \text{ or } ST^m.$$

Since  $S^{-1} = S$  and  $(T^n)^{-1} = T^{-n}$ , it follows that

$$g = T^{m_1} S \dots T^{m_k} S T^m \text{ or } g = T^{m_1} S \dots T^{m_k} T^m. \quad \square$$

The proof of Proposition 12 may be regarded as an analogue of the continued fraction algorithm, since

$$T^{m_1} S \dots T^{m_k} S T^m z = m_1 - \frac{1}{m_2 - \frac{1}{\ddots - \frac{1}{m_k - \frac{1}{m + z}}}.$$

Obviously  $\Gamma$  is also generated by  $S$  and  $R := ST$ . The transformation  $R$  has order 3, since

$$R(z) = -1/(z + 1), \quad R^2(z) = -(z + 1)/z, \quad R^3(z) = z.$$

We are going to show that all other relations between the generators  $S$  and  $R$  are consequences of the relations  $S^2 = R^3 = I$ , so that  $\Gamma$  is the *free product* of a cyclic group of order 2 and a cyclic group of order 3.

Partition the upper half-plane  $\mathcal{H}$  by putting

$$A = \{z \in \mathcal{H} : \Re z < 0\}, \quad B = \{z \in \mathcal{H} : \Re z \geq 0\}.$$

It is easily verified that

$$SA \subset B, \quad RB \subset A, \quad R^2B \subset A$$

(where the inclusions are strict). If  $g' = SR^{\varepsilon_1} SR^{\varepsilon_2} \dots SR^{\varepsilon_n}$  for some  $n \geq 1$ , where  $\varepsilon_j \in \{1, 2\}$ , it follows that  $g'B \subset B$  and  $g'SA \subset B$ . Similarly, if  $g'' = R^{\varepsilon_1} S \dots R^{\varepsilon_n}$ , then  $g''B \subset A$  and  $g''SA \subset A$ . By taking account of the relations  $S^2 = R^3 = I$ , every  $g \in \Gamma$  can be written in one of the forms

$$I, \quad S, \quad g', \quad g'', \quad g'S, \quad g''S.$$

But, by what has just been said, no element except the first is the identity transformation.

The modular group is *discrete*, since there exists a neighbourhood of the identity transformation which contains no other element of  $\Gamma$ .

**Proposition 13** *The open set*

$$F = \{z \in \mathcal{H} : -1/2 < \Re z < 1/2, |z| > 1\}$$

(see Figure 1) is a fundamental domain for the modular group  $\Gamma$ , i.e. distinct points of  $F$  are not equivalent and each point of  $\mathcal{H}$  is equivalent to some point of  $F$  or its boundary  $\partial F$ .

*Proof* For any  $z \in \mathbb{C}$  we write  $z = x + iy$ , where  $x, y \in \mathbb{R}$ . We show first that no two points of  $F$  are equivalent. Assume on the contrary that there exist distinct points  $z, z' \in F$  with  $y' \geq y$  such that

$$z' = (az + b)/(cz + d)$$

for some  $a, b, c, d \in \mathbb{Z}$  with  $ad - bc = 1$ . If  $c = 0$ , then  $a = d = \pm 1$ ,  $b \neq 0$  and  $z' = z + b/d$ , which is impossible for  $z, z' \in F$ . Hence  $c \neq 0$ . Since

$$y' = y/|cz + d|^2,$$

we have  $|cz + d| \leq 1$ . Thus  $|z + d/c| \leq 1/|c|$ , which is impossible not only if  $|c| \geq 2$  but also if  $c = \pm 1$ .

We now show that any  $z_0 \in \mathcal{H}$  is equivalent to a point of the closure  $\bar{F} = F \cup \partial F$ . We can choose  $m_0 \in \mathbb{Z}$  so that  $z_1 = z_0 + m_0$  satisfies  $|x_1| \leq 1/2$ . If  $|z_1| \geq 1$ , there is nothing more to do. Thus we now suppose  $|z_1| < 1$ . Put  $z_2 = -1/\bar{z}_1$ . Then

$$y_2 = y_1/|z_1|^2 > y_1$$

and actually  $y_2 \geq 2y_1$  if  $y_1 \leq 1/2$ , since then  $|z_1|^2 \leq 1/4 + 1/4 = 1/2$ . We now repeat the process, with  $z_2$  in place of  $z_0$ , and choose  $m_2 \in \mathbb{Z}$  so that  $z_3 = z_2 + m_2$  satisfies  $|x_3| \leq 1/2$ . From  $z_3 = (m_2 z_1 - 1)/z_1$  we obtain

$$|z_3|^2 = \{(m_2 x_1 - 1)^2 + (m_2 y_1)^2\}/(x_1^2 + y_1^2).$$

Assume  $|z_3| < 1$ . Then  $m_2 \neq 0$  and also  $m_2 \neq \pm 1$ , since  $|1 \pm x_1| \geq 1/2 \geq |x_1|$ . If  $|m_2| \geq 2$ , then  $|z_3|^2 \geq 4|y_1|^2$  and hence  $y_1 < 1/2$ . Thus in passing from  $z_1$  to  $z_3$  we obtain either  $z_3 \in \bar{F}$  or  $y_3 = y_2 \geq 2y_1$ . Hence, after repeating the process finitely many times we must obtain a point  $z_{2k+1} \in \bar{F}$ .  $\square$

Proposition 13 implies that the sets  $\{g(\bar{F}) : g \in \Gamma\}$  form a tiling of  $\mathcal{H}$ , since

$$\mathcal{H} = \bigcup_{g \in \Gamma} g(\bar{F}), \quad g(F) \cap g'(F) = \emptyset \text{ if } g, g' \in \Gamma \text{ and } g \neq g'.$$

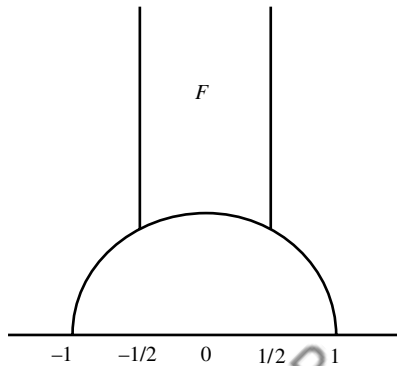


Fig. 1. Fundamental domain for  $\Gamma$ .

This is illustrated in Figure 2, where the domain  $g(F)$  is represented simply by the group element  $g$ .

There is an interesting connection between the modular group and binary quadratic forms. The *discriminant* of a binary quadratic form

$$f = ax^2 + bxy + cy^2$$

with coefficients  $a, b, c \in \mathbb{R}$  is  $D := b^2 - 4ac$ . The quadratic form is *indefinite* (i.e. assumes both positive and negative values) if and only if  $D > 0$ , and *positive definite* (i.e. assumes only positive values unless  $x = y = 0$ ) if and only if  $D < 0$ ,  $a > 0$ , which implies also  $c > 0$ . (If  $D = 0$ , the quadratic form is proportional to the square of a linear form.)

If we make a linear change of variables

$$x = \alpha x' + \beta y', \quad y = \gamma x' + \delta y',$$

where  $\alpha, \beta, \gamma, \delta \in \mathbb{Z}$  and  $\alpha\delta - \beta\gamma = 1$ , the quadratic form  $f$  is transformed into the quadratic form

$$f' = a'x'^2 + b'x'y' + c'y'^2,$$

where

$$\begin{aligned} a' &= a\alpha^2 + b\alpha\gamma + c\gamma^2, \\ b' &= 2a\alpha\beta + b(\alpha\delta + \beta\gamma) + 2c\gamma\delta, \\ c' &= a\beta^2 + b\beta\delta + c\delta^2, \end{aligned}$$

and hence

$$b'^2 - 4a'c' = b^2 - 4ac = D.$$

The quadratic forms  $f$  and  $f'$  are said to be *properly equivalent*.

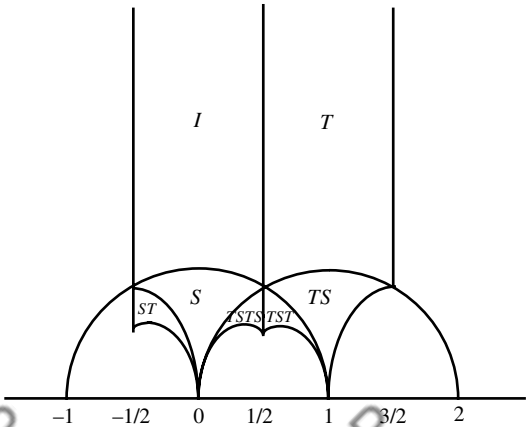


Fig. 2. Tiling of  $\mathcal{H}$  by  $T$ .

Thus properly equivalent forms have the same discriminant. As the name implies, proper equivalence is indeed an equivalence relation. Moreover, any form properly equivalent to an indefinite form is again indefinite, and any form properly equivalent to a positive definite form is again positive definite.

We will now show that any binary quadratic form is properly equivalent to one which is in some sense canonical. The indefinite and positive definite cases will be treated separately.

Suppose first that  $f$  is positive definite, so that  $D < 0$ ,  $a > 0$  and  $c > 0$ . With the quadratic form  $f$  we associate a point  $\tau(f)$  of the upper half-plane  $\mathcal{H}$ , namely

$$\tau(f) = (-b + i\sqrt{-D})/2a.$$

Thus  $\tau(f)$  is the root with positive imaginary part of the polynomial  $at^2 + bt + c$ . Conversely, for any given  $D < 0$  and  $\tau \in \mathcal{H}$ , there is a unique positive definite quadratic form  $f$  with discriminant  $D$  such that  $\tau(f) = \tau$ . In fact, if  $\tau = \zeta + i\eta$ , where  $\zeta, \eta \in \mathbb{R}$  and  $\eta > 0$ , we must take

$$a = \sqrt{(-D)}/2\eta, \quad b = -2a\zeta, \quad c = (b^2 - D)/4a.$$

Let  $f'$ , as above, be a form properly equivalent to  $f$ . If  $t = (\alpha t' + \beta)/(\gamma t' + \delta)$ , then

$$at^2 + bt + c = (a't'^2 + b't' + c')/(\gamma t' + \delta)^2.$$

It follows that if  $\tau = \tau(f)$  and  $\tau' = \tau(f')$ , then  $\tau = (\alpha\tau' + \beta)/(\gamma\tau' + \delta)$ . Thus  $\tau'$  is properly equivalent to  $\tau$ , in the terminology introduced in Section 1.

By Proposition 13 we may choose the change of variables so that  $\tau' \in \bar{F}$ , i.e.

$$-1/2 \leq \Re \tau' \leq 1/2, \quad |\tau'| \geq 1.$$

It is easily verified that this is the case if and only if for  $f'$  we have

$$|b'| \leq a', \quad 0 < a' \leq c'.$$

Such a quadratic form  $f'$  is said to be *reduced*. Thus every positive definite binary quadratic form is properly equivalent to a reduced form. (It is possible to ensure that every positive definite binary quadratic form is properly equivalent to a unique reduced form by slightly restricting the definition of ‘reduced’, but we will have no need of this.)

If the coefficients of  $f$  are integers, then so also are the coefficients of  $f'$  and  $\tau, \tau'$  are complex quadratic irrationals. There are only finitely many reduced forms  $f$  with integer coefficients and with a given discriminant  $D < 0$ . For, if  $f$  is reduced, then

$$4b^2 \leq 4a^2 \leq 4ac = b^2 - D$$

and hence  $b^2 \leq -D/3$ . Since  $4ac = b^2 - D$ , for each of the finitely many possible values of  $b$  there are only finitely many possible values for  $a$  and  $c$ .

A quadratic form  $f = ax^2 + bxy + cy^2$  is said to be *primitive* if the coefficients  $a, b, c$  are integers with greatest common divisor 1. For any integer  $D < 0$ , let  $h^\dagger(D)$

denote the number of primitive positive definite quadratic forms with discriminant  $D$  which are properly inequivalent. By what has been said,  $h^\dagger(D)$  is finite.

Consider next the indefinite case:

$$f = ax^2 + bxy + cy^2$$

where  $a, b, c \in \mathbb{R}$  and  $D > 0$ . If  $a \neq 0$ , we can write

$$f = a(x - \xi y)(x - \eta y),$$

where  $\xi, \eta$  are the distinct real roots of the polynomial  $at^2 + bt + c$ . It follows from Lemma 0 that, if  $\xi$  and  $\eta$  are irrational, then  $f$  is properly equivalent to a form  $f'$  for which  $\xi' > 1$  and  $-1 < \eta' < 0$ . Such a quadratic form  $f'$  is said to be *reduced*. Evidently  $f'$  is reduced if and only if  $-f'$  is reduced. Thus we may suppose  $a' > 0$ , and then  $f'$  is reduced if and only if

$$0 < \sqrt{D} + b' < 2a' < \sqrt{D} - b'.$$

If the coefficients of  $f$  are integers and the positive integer  $D$  is not a square, then  $a \neq 0$  and  $\xi, \eta$  are conjugate real quadratic irrationals. In this case, as we already saw in Section 3, there are only finitely many reduced forms with discriminant  $D$ . For any integer  $D > 0$  which is not a square, let  $h^\dagger(D)$  denote the number of primitive quadratic forms with discriminant  $D$  which are properly inequivalent. By what has been said,  $h^\dagger(D)$  is finite.

It should be noted that, for any quadratic form  $f$  with integer coefficients, the discriminant  $D \equiv 0$  or  $1 \pmod{4}$ . Moreover, for any  $D \equiv 0$  or  $1 \pmod{4}$ , there is a quadratic form  $f$  with integer coefficients and with discriminant  $D$ ; for example,

$$\begin{aligned} f &= x^2 - Dy^2/4 & \text{if } D \equiv 0 \pmod{4}, \\ f &= x^2 + xy + (1-D)y^2/4 & \text{if } D \equiv 1 \pmod{4}. \end{aligned}$$

The preceding results for quadratic forms can also be restated in terms of quadratic fields. By making correspond to the ideal with basis  $\beta = a, \gamma = b + c\omega$  in the quadratic field  $\mathbb{Q}(\sqrt{d})$  the binary quadratic form

$$\{\beta\beta'x^2 + (\beta\gamma' + \beta'\gamma)xy + \gamma\gamma'y^2\}/ac,$$

one can establish a bijective map between 'strict' equivalence classes of ideals in  $\mathbb{Q}(\sqrt{d})$  and proper equivalence classes of binary quadratic forms with discriminant  $D$ , where

$$\begin{aligned} D &= 4d & \text{if } d \equiv 2 \text{ or } 3 \pmod{4}, \\ D &= d & \text{if } d \equiv 1 \pmod{4}. \end{aligned}$$

(The middle coefficient  $b$  of  $f = ax^2 + bxy + cy^2$  was not required to be even in order to obtain this one-to-one correspondence.) Since any ideal class is either a strict ideal class or the union of two strict ideal classes, the finiteness of the class number  $h(d)$  of the quadratic field  $\mathbb{Q}(\sqrt{d})$  thus follows from the finiteness of  $h^\dagger(D)$ .

## 6 Non-Euclidean Geometry

There is an important connection between the modular group and the non-Euclidean geometry of Bolyai (1832) and Lobachevski (1829). It was first pointed out by Beltrami (1868) that their *hyperbolic geometry* is the geometry on a manifold of constant curvature. In the model of Poincaré (1882) for two-dimensional hyperbolic geometry the underlying space is taken to be the upper half-plane  $\mathcal{H}$ . A ‘line’ is either a semi-circle with centre on the real axis or a half-line perpendicular to the real axis. It follows that through any two distinct points there passes exactly one ‘line’. However, through a given point not on a given ‘line’ there passes more than one ‘line’ having no point in common with the given ‘line’.

Although Euclid’s parallel axiom fails to hold, all the other axioms of Euclidean geometry are satisfied. Poincaré’s model shows that if Euclidean geometry is free from contradiction, then so also is hyperbolic geometry. Before the advent of non-Euclidean geometry there had been absolute faith in Euclidean geometry. It is realized today that it is a matter for experiment to determine what kind of geometry best describes our physical world.

Poincaré’s model will now be examined in more detail (with the constant curvature normalized to have the value  $-1$ ). A curve  $\gamma$  in  $\mathcal{H}$  is specified by a continuously differentiable function  $z(t) = x(t) + iy(t)$  ( $a \leq t \leq b$ ). The (hyperbolic) *length* of  $\gamma$  is defined to be

$$\ell(\gamma) = \int_a^b y(t)^{-1} |dz/dt| dt.$$

It follows from this definition that the ‘line’ segment joining two points  $z, w$  of  $\mathcal{H}$  has length

$$d(z, w) = \ln \frac{|z - \bar{w}| + |z - w|}{|z - \bar{w}| - |z - w|}.$$

It may be shown that any other curve joining  $z$  and  $w$  has greater length. Thus the ‘lines’ are *geodesics*.

For any  $z_0 \in \mathcal{H}$ , there is a unique geodesic through  $z_0$  in any specified direction. Also, for any distinct real numbers  $\xi, \eta$ , there is a unique geodesic which intersects the real axis at  $\xi, \eta$ , namely the semicircle with centre at  $(\xi + \eta)/2$ . (By abuse of language we say ‘ $\xi$ ’, for example, when we mean the point  $(\xi, 0)$ .)

A linear fractional transformation

$$z' = f(z) = (az + b)/(cz + d),$$

where  $a, b, c, d \in \mathbb{R}$  and  $ad - bc = 1$ , maps the upper half-plane  $\mathcal{H}$  onto itself and maps ‘lines’ onto ‘lines’. Moreover, if the curve  $\gamma$  is mapped onto the curve  $\gamma'$ , then  $\ell(\gamma) = \ell(\gamma')$ , since  $\mathcal{I}f(z) = \mathcal{I}z/|cz + d|^2$  and  $df/dz = 1/|cz + d|^2$ . In particular,

$$d(z, w) = d(z', w').$$

Thus a linear fractional transformation of the above form is an *isometry*. It may be shown that any isometry is either a linear fractional transformation of this form or is

obtained by composing such a transformation with the (orientation-reversing) transformation  $x + iy \rightarrow -x + iy$ . For any two ‘lines’  $L$  and  $L'$ , there is an isometry which maps  $L$  onto  $L'$ .

We may define *angles* to be the same as in Euclidean geometry, since any linear fractional transformation is conformal. The (hyperbolic) *area* of a domain  $D \subset \mathcal{H}$ , defined by

$$\mu(D) = \iint_D y^{-2} dx dy,$$

is invariant under any isometry. In particular, this gives  $\pi - (\alpha + \beta + \gamma)$  for the area of a ‘triangle’ with angles  $\alpha, \beta, \gamma$ . Since the angles are non-negative, the area of a ‘triangle’ is at most  $\pi$  and, since the area is necessarily positive, the sum of the angles of a ‘triangle’ is less than  $\pi$ .

For example, if  $F$  is the fundamental domain of the modular group  $\Gamma$ , then  $\bar{F}$  is a ‘triangle’ with angles  $\pi/3, \pi/3, 0$  and hence the area of  $\bar{F}$  is  $\pi - 2\pi/3 = \pi/3$ . For any fixed  $z_0 \in F$  on the imaginary axis, we may characterize  $F$  as the set of all  $z \in \mathcal{H}$  such that, for every  $g \in \Gamma$  with  $g \neq I$ ,

$$d(z, z_0) < d(z, g(z_0)) = d(g^{-1}(z), z_0).$$

By identifying two points  $z, z'$  of  $\mathcal{H}$  if  $z' = g(z)$  for some  $g \in \Gamma$  we obtain the *quotient space*  $\mathcal{M} = \mathcal{H}/\Gamma$ . Equivalently, we may regard  $\mathcal{M}$  as the closure  $\bar{F}$  of the fundamental domain  $F$  with the boundary point  $-1/2 + iy$  identified with the boundary point  $1/2 + iy$  ( $1 \leq y < \infty$ ) and the boundary point  $-e^{-i\theta}$  identified with the boundary point  $e^{i\theta}$  ( $0 < \theta < \pi/2$ ).

Since the elements of  $\Gamma$  are isometries of  $\mathcal{H}$ , the metric on  $\mathcal{H}$  induces a metric on  $\mathcal{M}$  in which the geodesics are the projections onto  $\mathcal{M}$  of the geodesics in  $\mathcal{H}$ . Thus if we regard  $\mathcal{M}$  as  $\bar{F}$  with appropriate boundary points identified, then a geodesic in  $\mathcal{M}$  will be a sequence of geodesic arcs in  $F$ , each with initial point and endpoint on the boundary of  $F$ , so that the initial point of one arc is the point identified to the endpoint of the preceding arc.

Let  $L$  be a geodesic in  $\mathcal{H}$  which intersects the real axis in irrational points  $\zeta, \eta$  such that  $\zeta > 1, -1 < \eta < 0$  and let

$$\zeta = [a_0, a_1, a_2, \dots], \quad -1/\eta = [a_{-1}, a_{-2}, \dots]$$

be the continued fraction expansions of  $\zeta$  and  $-1/\eta$ . If we choose  $\zeta$  and  $\eta = \zeta'$  to be conjugate quadratic irrationals then, by Proposition 7, the doubly-infinite sequence

$$[\dots, a_{-2}, a_{-1}, a_0, a_1, a_2, \dots]$$

is periodic and it is not difficult to see that the geodesic in  $\mathcal{M}$  obtained by projection from  $L$  is closed. Artin (1924) showed that there are other geodesics which behave very differently. Let the convergents of  $\zeta$  be  $p_n/q_n$  and put

$$\zeta = (p_{n-1}\zeta_n + p_{n-2})/(q_{n-1}\zeta_n + q_{n-2}), \quad \eta = (p_{n-1}\eta_n + p_{n-2})/(q_{n-1}\eta_n + q_{n-2}).$$

Then

$$\xi_n = [a_n, a_{n+1}, \dots], \quad -1/\eta_n = [a_{n-1}, a_{n-2}, \dots],$$

and  $\xi_n > 1$ ,  $-1 < \eta_n < 0$ . Moreover, if  $n$  is even, then  $\xi$  and  $\eta$  are properly equivalent to  $\xi_n$  and  $\eta_n$  respectively. If we choose  $\xi$  so that the sequence  $a_0, a_1, a_2, \dots$  contains each finite sequence of positive integers (and hence contains it infinitely often), then the corresponding geodesic in  $\mathcal{M}$  passes arbitrarily close to every point of  $\mathcal{M}$  and to every direction at that point.

Some much-studied subgroups of the modular group are the *congruence subgroups*  $\Gamma(n)$ , consisting of all linear fractional transformations  $z \rightarrow (az + b)/(cz + d)$  in  $\Gamma$  congruent to the identity transformation, i.e.

$$a \equiv d \equiv \pm 1, \quad b \equiv c \equiv 0 \pmod{n}.$$

We may in the same way investigate the geodesics in the *quotient space*  $\mathcal{H}/\Gamma(n)$ . In the case  $n = 3$  it has been shown by Lehner and Sheingorn (1984) that there is an interesting connection with the *Markov spectrum*.

In Section 2 we defined, for any irrational number  $\xi$  with convergents  $p_n/q_n$ ,

$$M(\xi) = \overline{\lim}_{n \rightarrow \infty} q_n^{-1} |q_n \xi - p_n|^{-1},$$

and we noted that  $M(\xi) = M(\eta)$  if  $\xi$  and  $\eta$  are equivalent. It is not difficult to show that there are uncountably many inequivalent  $\xi$  for which  $M(\xi) = 3$ . However, it was shown by Markov (1879/80) that there is a sequence of real quadratic irrationals  $\xi^{(k)}$  such that  $M(\xi) < 3$  if and only if  $\xi$  is equivalent to  $\xi^{(k)}$  for some  $k$ . If  $\mu_k = M(\xi^{(k)})$ , then  $\mu_1 < \mu_2 < \mu_3 < \dots$  and  $\mu_k \rightarrow 3$  as  $k \rightarrow \infty$ . Although  $\mu_k$  is irrational,  $\mu_k^2$  is rational. The first few values are

$$\begin{aligned} \mu_1 &= 5^{1/2} = 2.236\dots, & \mu_2 &= 8^{1/2} = 2.828\dots, \\ \mu_3 &= (221)^{1/2}/5 = 2.973\dots, & \mu_4 &= (1517)^{1/2}/13 = 2.996\dots \end{aligned}$$

As we already showed in Section 2, we can take  $\xi^{(1)} = (1 + \sqrt{5})/2$  and  $\xi^{(2)} = 1 + \sqrt{2}$ .

Lehner and Sheingorn showed that the simple closed geodesics in  $\mathcal{H}/\Gamma(3)$  are just the projections of the geodesics in  $\mathcal{H}$  whose endpoints  $\xi, \eta$  on the real axis are conjugate quadratic irrationals equivalent to  $\xi^{(k)}$  for some  $k$ .

There is a recursive procedure for calculating the quantities  $\mu_k$  and  $\xi^{(k)}$ . A *Markov triple* is a triple  $(u, v, w)$  of positive integers such that

$$u^2 + v^2 + w^2 = 3uvw.$$

If  $(u, v, w)$  is a Markov triple, then so also are  $(3uw - v, u, w)$  and  $(3uv - w, u, v)$ . They are distinct from the original triple if  $u = \max(u, v, w)$ , since then  $u < 3uw - v$  and  $u < 3uv - w$ . They are also distinct from one another if  $w < v$ . Starting from the trivial triple  $(1, 1, 1)$ , all Markov triples can be obtained by repeated applications of this process. The successive values of  $u = \max(u, v, w)$  are 1, 2, 5, 13, 29, ... The numbers  $\mu_k$  and  $\xi^{(k)}$  are the corresponding successive values of  $(9 - 4/u^2)^{1/2}$  and  $(9 - 4/u^2)^{1/2}/2 + 1/2 + v/uw$ .

It was conjectured by Frobenius (1913) that a Markov triple is uniquely determined by its greatest element. This has been verified whenever the greatest element does not exceed  $10^{140}$ . It has also been proved when the greatest element is a prime (and in some other cases) by Baragar (1996), using the theory of quadratic fields.

## 7 Complements

There is an important analogue of the continued fraction algorithm for infinite series. Let  $K$  be an arbitrary field and let  $F$  denote the set of all formal Laurent series

$$f = \sum_{n \in \mathbb{Z}} \alpha_n t^n$$

with coefficients  $\alpha_n \in K$  such that  $\alpha_n \neq 0$  for at most finitely many  $n > 0$ . If

$$g = \sum_{n \in \mathbb{Z}} \beta_n t^n$$

is also an element of  $F$ , and if we define addition and multiplication by

$$f + g = \sum_{n \in \mathbb{Z}} (\alpha_n + \beta_n) t^n, \quad fg = \sum_{n \in \mathbb{Z}} \gamma_n t^n,$$

where  $\gamma_n = \sum_{j+k=n} \alpha_j \beta_k$ , then  $F$  acquires the structure of a commutative ring. In fact,  $F$  is a field. For, if  $f = \sum_{n \leq v} \alpha_n t^n$ , where  $\alpha_v \neq 0$ , we obtain  $g = \sum_{n \leq -v} \beta_n t^n$  such that  $fg = 1$  by solving successively the equations

$$\begin{aligned} \alpha_v \beta_{-v} &= 1 \\ \alpha_v \beta_{-v-1} + \alpha_{v-1} \beta_{-v} &= 0 \\ \alpha_v \beta_{-v-2} + \alpha_{v-1} \beta_{-v-1} + \alpha_{v-2} \beta_{-v} &= 0 \\ &\dots \end{aligned}$$

Define the absolute value of an element  $f = \sum_{n \in \mathbb{Z}} \alpha_n t^n$  of  $F$  by putting

$$|O| = 0, \quad |f| = 2^{v(f)} \quad \text{if } f \neq O,$$

where  $v(f)$  is the greatest integer  $n$  such that  $\alpha_n \neq 0$ . It is easily verified that

$$|fg| = |f||g|, \quad |f + g| \leq \max(|f|, |g|),$$

and  $|f + g| = \max(|f|, |g|)$  if  $|f| \neq |g|$ .

For any  $f = \sum_{n \in \mathbb{Z}} \alpha_n t^n \in F$ , let

$$\lfloor f \rfloor = \sum_{n \geq 0} \alpha_n t^n, \quad \{f\} = \sum_{n < 0} \alpha_n t^n$$

denote respectively its polynomial and strictly proper parts. Then  $|\{f\}| < 1$ , and  $|\lfloor f \rfloor| = |f|$  if  $|f| \geq 1$ , i.e. if  $\lfloor f \rfloor \neq O$ .

If  $f_0 := f$  is not the formal Laurent series of a rational function, we can write

$$f_0 = a_0 + 1/f_1,$$

where  $a_0 = \lfloor f_0 \rfloor$  and  $|f_1| > 1$ . In the same way,

$$f_1 = a_1 + 1/f_2,$$

where  $a_1 = \lfloor f_1 \rfloor$  and  $|f_2| > 1$ . Continuing in this way, we obtain the *continued fraction expansion*  $[a_0, a_1, a_2, \dots]$  of  $f$ . In the same way as for real numbers, if we define polynomials  $p_n, q_n$  by the recurrence relations

$$p_n = a_n p_{n-1} + p_{n-2}, \quad q_n = a_n q_{n-1} + q_{n-2} \quad (n \geq 0),$$

with  $p_{-2} = q_{-1} = 0, p_{-1} = q_{-2} = 1$ , then

$$p_n q_{n-1} - p_{n-1} q_n = (-1)^{n+1} \quad (n \geq 0),$$

$$f = (p_n f_{n+1} + p_{n-1}) / (q_n f_{n+1} + q_{n-1}) \quad (n \geq 0),$$

and so on. In addition, however, we now have

$$|a_n| = |f_n| > 1 \quad (n \geq 1),$$

from which we obtain by induction

$$|p_n| = |a_n| |p_{n-1}| > |p_{n-1}|, \quad |q_n| = |a_n| |q_{n-1}| > |q_{n-1}| \quad (n \geq 1).$$

Hence

$$|p_n| = |a_0 a_1 \cdots a_n|, \quad |q_n| = |a_1 \cdots a_n| \quad (n \geq 1).$$

From the relation  $q_n f - p_n = (-1)^n / (q_n f_{n+1} + q_{n-1})$  we further obtain

$$|q_n f - p_n| = |q_{n+1}|^{-1},$$

since

$$|q_n f_{n+1} + q_{n-1}| = |q_n f_{n+1}| = |q_n| |a_{n+1}| = |q_{n+1}|.$$

In particular,  $|q_n f - p_n| < 1$  and hence

$$p_n = \lfloor q_n f \rfloor, \quad \{q_n f\} = |q_{n+1}|^{-1} \quad (n \geq 1).$$

Thus  $p_n$  is readily determined from  $q_n$ . Furthermore,

$$|f - p_n/q_n| = |q_n|^{-1} |q_{n+1}|^{-1} \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

The rational function  $p_n/q_n$  is called the *n-th convergent* of  $f$ . The polynomials  $a_n$  are called the *partial quotients*, and the Laurent series  $f_n$  the *complete quotients*, in the continued fraction expansion of  $f$ .

The continued fraction algorithm can also be applied when  $f$  is the formal Laurent expansion of a rational function, but in this case the process terminates after a finite number of steps. If  $a_0, a_1, a_2, \dots$  is any finite or infinite sequence of polynomials with  $|a_n| > 1$  for  $n \geq 1$ , there is a unique formal Laurent series  $f$  with  $[a_0, a_1, a_2, \dots]$  as its continued fraction expansion.

For formal Laurent series there are sharper Diophantine properties than for real numbers:

**Proposition 14** *Let  $f$  be a formal Laurent series with convergents  $p_n/q_n$  and let  $p, q$  be polynomials with  $q \neq 0$ .*

(i) *If  $|q| < |q_{n+1}|$  and  $p/q \neq p_n/q_n$ , then*

$$|qf - p| \geq |q_{n-1}f - p_{n-1}| = |q_n|^{-1}.$$

(ii) *If  $|qf - p| < |q|^{-1}$ , then  $p/q$  is a convergent of  $f$ .*

*Proof* (i) Assume on the contrary that  $|qf - p| < |q_n|^{-1}$ . Since

$$q_n(qf - p) - q(q_n f - p_n) = qp_n - pq_n \neq 0$$

and  $|q_n||qf - p| < 1$ , we must have

$$|q||q_{n+1}|^{-1} = |q||q_n f - p_n| = |qp_n - pq_n| \geq 1,$$

which is contrary to hypothesis.

(ii) Assume that  $p/q$  is not a convergent of  $f$ . If  $f = p_N/q_N$  is a rational function then  $|q| < |q_N|$ , since

$$1 \leq |qp_N - pq_N| = |qf - p||q_N| < |q|^{-1}|q_N|.$$

Thus, whether or not  $f$  is rational, we can choose  $n$  so that  $|q_n| \leq |q| < |q_{n+1}|$ . Hence, by (i),

$$|qf - p| \geq |q_n|^{-1} \geq |q|^{-1},$$

which is a contradiction. □

It was shown by Abel (1826) that, for any complex polynomial  $D(t)$  which is not a square, the ‘Pell’ equation  $X^2 - D(t)Y^2 = 1$  has a solution in polynomials  $X(t), Y(t)$  of positive degree if and only if  $\sqrt{D(t)}$  may be represented as a periodic continued fraction:  $\sqrt{D(t)} = [a_0, \overline{a_1, \dots, a_h}]$ , where  $a_h = 2a_0$  and  $a_i = a_{h-i}$  ( $i = 1, \dots, h-1$ ) are polynomials of positive degree. By differentiation one obtains

$$XX'/Y = Y'D + (1/2)YD'.$$

It follows that  $Y$  divides  $X'$ , since  $X$  and  $Y$  are relatively prime, and

$$(X + Y\sqrt{D})' = (X + Y\sqrt{D})X'/Y\sqrt{D}.$$

Thus the ‘abelian’ integral

$$\int X'(t)dt/Y(t)\sqrt{D(t)}$$

is actually the elementary function  $\log\{X(t) + Y(t)\sqrt{D(t)}\}$ .

Some remarkable results have recently been obtained on the approximation of algebraic numbers by rational numbers, which deserve to be mentioned here, even though the proofs are beyond our scope.

A complex number  $\zeta$  is said to be an *algebraic number*, or simply *algebraic*, of degree  $d$  if it is a root of a polynomial of degree  $d$  with rational coefficients which is irreducible over the rational field  $\mathbb{Q}$ . Thus an algebraic number of degree 2 is just a quadratic irrational.

For any irrational number  $\zeta$ , there exist infinitely many rational numbers  $p/q$  such that

$$|\zeta - p/q| < 1/q^2,$$

since the inequality is satisfied by any convergent of  $\zeta$ . It was shown by Roth (1955) that if  $\zeta$  is a real algebraic number of degree  $d \geq 2$  then, for any given  $\varepsilon > 0$ , there exist only finitely many rational numbers  $p/q$  with  $q > 0$  such that

$$|\zeta - p/q| < 1/q^{2+\varepsilon}.$$

The proof does not provide a bound for the magnitude of the rational numbers which satisfy the inequality, but it does provide a bound for their number. Roth's result was the culmination of a line of research that was begun by Thue (1909), and further developed by Siegel (1921) and Dyson (1947).

A sharpening of Roth's result has been *conjectured* by Lang (1965): if  $\zeta$  is a real algebraic number of degree  $d \geq 2$  then, for any given  $\varepsilon > 0$ , there exist only finitely many rational numbers  $p/q$  with  $q > 1$  such that

$$|\zeta - p/q| < 1/q^2 (\log q)^{1+\varepsilon}.$$

An even stronger sharpening has been conjectured by P.M. Wong (1989) in which  $(\log q)^{1+\varepsilon}$  is replaced by  $(\log q)(\log \log q)^{1+\varepsilon}$  with  $q > 2$ .

For real algebraic numbers of degree 2 we already know more than this. For, if  $\zeta$  is a real quadratic irrational, its partial quotients are bounded and so there exists a constant  $c = c(\zeta) > 0$  such that  $|\zeta - p/q| > c/q^2$  for every rational number  $p/q$ . It is a long-standing conjecture that this is false for any real algebraic number  $\zeta$  of degree  $d > 2$ .

It is not difficult to show that Roth's theorem may be restated in the following homogeneous form: if

$$L_1(u, v) = \alpha u + \beta v, \quad L_2(u, v) = \gamma u + \delta v,$$

are linearly independent linear forms with algebraic coefficients  $\alpha, \beta, \gamma, \delta$ , then, for any given  $\varepsilon > 0$ , there exist at most finitely many integers  $x, y$ , not both zero, such that

$$|L_1(x, y)L_2(x, y)| < \max(|x|, |y|)^{-\varepsilon}.$$

The *subspace theorem* of W. Schmidt (1972) generalizes Roth's theorem in this form to higher dimensions. In the stronger form given it by Vojta (1989) it says: if  $L_1(\mathbf{u}), \dots, L_n(\mathbf{u})$  are linearly independent linear forms in  $n$  variables  $\mathbf{u} = (u_1, \dots, u_n)$  with (real or complex) algebraic coefficients, then there exist finitely many proper linear subspaces  $V_1, \dots, V_h$  of  $\mathbb{Q}^n$  such that every nonzero  $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{Z}^n$  for which

$$|L_1(\mathbf{x}) \cdots L_n(\mathbf{x})| < \|\mathbf{x}\|^{-\varepsilon},$$

where  $\|x\| = \max(|x_1|, \dots, |x_n|)$ , is contained in some subspace  $V_i$ , except for finitely many points whose number may depend on  $\varepsilon$ . A new proof of Schmidt's subspace theorem has been given by Faltings and Wüstholz (1994). The subspace theorem has also been given a more quantitative form by Schmidt (1989) and Evertse (1996). These results have immediate applications to the simultaneous approximation of several algebraic numbers.

Vojta (1987) has developed a remarkable analogy between the approximation of algebraic numbers by rationals and the theory of Nevanlinna (1925) on the value distribution of meromorphic functions, in which Roth's theorem corresponds to Nevanlinna's second main theorem. Although the analogy is largely formal, it is suggestive in both directions. It has already led to new proofs for the theorems of Roth and Schmidt, and to a proof of the Mordell conjecture (discussed below) which is quite different from the original proof by Faltings.

Roth's theorem has an interesting application to Diophantine equations. Let

$$f(z) = a_0 z^n + a_1 z^{n-1} + \dots + a_n$$

be a polynomial of degree  $n \geq 3$  with integer coefficients whose roots are distinct and not rational. Let

$$f(u, v) = a_0 u^n + a_1 u^{n-1} v + \dots + a_n v^n$$

be the corresponding homogeneous polynomial and let  $g(u, v)$  be a polynomial of degree  $m \geq 0$  with integer coefficients. We will deduce from Roth's theorem that the equation

$$f(x, y) = g(x, y)$$

has at most finitely many solutions in integers if  $m \leq n - 3$ . This was already proved by Thue for  $m = 0$ .

Assume on the contrary that there exist infinitely many solutions in integers. Without loss of generality we may assume that there exist infinitely many integer solutions  $x, y$  for which  $|x| \leq |y|$ . Then there exists a constant  $c_1 > 0$  such that

$$|g(x, y)| \leq c_1 |y|^m.$$

Over the complex field  $\mathbb{C}$  the homogeneous polynomial  $f(u, v)$  has a factorization

$$f(u, v) = a_0 \prod_{j=1}^n (u - \zeta_j v),$$

where  $\zeta_1, \dots, \zeta_n$  are distinct algebraic numbers which are not rational. For at least one  $j$  we must have, for infinitely many  $x, y$ ,

$$|a_0| |x - \zeta_j y|^n \leq c_1 |y|^m$$

and hence

$$|x - \zeta_j y| \leq c_2 |y|^{m/n},$$

where  $c_2 = (c_1/|a_0|)^{1/n}$ . If  $k \neq j$ , then

$$\begin{aligned} |x - \zeta_k y| &\geq |(\zeta_j - \zeta_k)y| - |x - \zeta_j y| \\ &\geq c_3|y| - c_2|y|^{m/n} \geq c_4|y|, \end{aligned}$$

where  $c_3, c_4$  are positive constants. It follows that

$$|a_0||x - \zeta_j y|c_4^{n-1}|y|^{n-1} \leq |f(x, y)| = |g(x, y)| \leq c_1|y|^m$$

and hence

$$|\zeta_j - x/y| \leq c_5/|y|^{n-m},$$

where the positive constant  $c_5$  depends only on the coefficients of  $f$  and  $g$ . Evidently this implies that  $\zeta_j$  is real. Since  $\zeta_j$  is not rational and  $m \leq n - 3$ , we now obtain a contradiction to Roth's theorem.

It is actually possible to characterize all polynomial Diophantine equations with infinitely many solutions. Let  $F(x, y)$  be a polynomial with rational coefficients which is irreducible over  $\mathbb{C}$ . It was shown by Siegel (1929), by combining his own results on the approximation of algebraic numbers with results of Mordell and Weil concerning the rational points on elliptic curves and Jacobian varieties, that if the equation

$$F(x, y) = 0 \tag{*}$$

has infinitely many integer solutions, then there exist polynomials or Laurent polynomials  $\phi(t), \psi(t)$  (not both constant) with coefficients from either the rational field  $\mathbb{Q}$  or a real quadratic field  $\mathbb{Q}(\sqrt{d})$ , where  $d > 0$  is a square-free integer, such that  $F(\phi(t), \psi(t))$  is identically zero. If  $\phi(t), \psi(t)$  are Laurent polynomials with coefficients from  $\mathbb{Q}(\sqrt{d})$ , they may be chosen to be invariant when  $t$  is replaced by  $t^{-1}$  and the coefficients are replaced by their conjugates in  $\mathbb{Q}(\sqrt{d})$ .

This implies, in particular, that the algebraic curve defined by (\*) may be transformed by a birational transformation with rational coefficients into either a linear equation  $ax + by + c = 0$  or a Pellian equation  $x^2 - dy^2 - m = 0$ . It is not significant that the birational transformation has rational, rather than integral, coefficients since, by combining a result of Mahler (1934) with the *Mordell conjecture*, it may be seen that the same conclusions hold if the equation (\*) has infinitely many solutions in rational numbers whose denominators involve only finitely many primes.

The conjecture of Mordell (1922) says that the equation (\*) has at most finitely many *rational* solutions if the algebraic curve defined by (\*) has genus  $g > 1$ . (The concept of *genus* will not be formally defined here, but we mention that the genus of an irreducible plane algebraic curve may be calculated by a procedure due to M. Noether.) The conjecture has now been proved by Faltings (1983), as will be mentioned in Chapter XIII. As mentioned also at the end of Chapter XIII, if the algebraic curve defined by (\*) has genus 1, then explicit bounds may be obtained for the number of integral points. It was already shown by Hilbert and Hurwitz (1890) that the algebraic curve defined by (\*) has genus 0 if and only if it is birationally equivalent over  $\mathbb{Q}$  either to a line or to a conic. There then exist rational functions  $\phi(t), \psi(t)$  (not both constant) with coefficients either from  $\mathbb{Q}$  or from a quadratic extension of  $\mathbb{Q}$  such that

$F(\phi(t), \psi(t))$  is identically zero. The coefficients may be taken from  $\mathbb{Q}$  if the curve has at least one non-singular rational point.

Thus in retrospect, and quite unfairly, Siegel's remarkable result may be seen as simply picking out those curves of genus 0 which have infinitely many integral points, a problem which had already been treated by Maillet (1919).

In this connection it may be mentioned that the formula for Pythagorean triples given in §5 of Chapter II may be derived from the parametrization of the unit circle  $x^2 + y^2 = 1$  by the rational functions

$$x(t) = (1 - t^2)/(1 + t^2), \quad y(t) = 2t/(1 + t^2).$$

## 8 Further Remarks

More extensive accounts of the theory of continued fractions are given in the books of Rockett and Szusz [45] and Perron [41]. Many historical references are given in Brezinski [12]. The first systematic account of the subject, which it is still a delight to read, was given in 1774 by Lagrange [32] in his additions to the French translation of Euler's *Algebra*.

The continued fraction algorithm is such a useful tool that there have been many attempts to generalize it to higher dimensions. Jacobi, in a paper published posthumously (1868), defined a continued fraction algorithm in  $\mathbb{R}^2$ . Perron (1907) extended his definition to  $\mathbb{R}^n$  and proved that convergence holds in the following weak sense: for a given nonzero  $\mathbf{x} \in \mathbb{R}^n$ , the Jacobi-Perron algorithm constructs recursively a sequence of bases  $\mathcal{B}^k = \{\mathbf{b}_1^k, \dots, \mathbf{b}_n^k\}$  of  $\mathbb{Z}^n$  such that, for each  $j \in \{1, \dots, n\}$ , the angle between the line  $O\mathbf{b}_j^k$  and the line  $O\mathbf{x}$  tends to zero as  $k \rightarrow \infty$ . More recently, other algorithms have been proposed for which convergence holds in the strong sense that, for each  $j \in \{1, \dots, n\}$ , the distance of  $\mathbf{b}_j^k$  from the line  $O\mathbf{x}$  tends to zero as  $k \rightarrow \infty$ . See Brentjes [11], Ferguson [22], Just [28] and Lagarias [31].

Proposition 2 was first proved by Serret [51]. Proposition 3 was proved by Lagrange. The complete characterization of best approximations is proved in the book of Perron.

Lambert (1766) proved that  $\pi$  was irrational by using a continued fraction expansion for  $\tan x$ . For the continued fraction expansion of  $\pi$ , see Choong *et al.* [15]. Badly approximable numbers are thoroughly surveyed by Shallit [52].

The theory of Diophantine approximation is treated more comprehensively in the books of Koksma [30], Cassels [13] and Schmidt [47].

The estimate  $O(\sqrt{D} \log D)$  for the period of the continued fraction expansion of a quadratic irrational with discriminant  $D$  is proved by elementary means in the book of Rockett and Szusz. Further references are given in Podsypanin [42].

The ancient Hindu method of solving Pell's equation is discussed in Selenius [49]. Tables for solving the Diophantine equation  $x^2 - dy^2 = m$ , where  $m^2 < d$ , are given in Patz [39]. Pell's equation plays a role in the negative solution of Hilbert's tenth problem, which asks for an algorithm to determine whether an arbitrary polynomial Diophantine equation is solvable in integers. See Davis *et al.* [18] and Jones and Matijasevic [26].

The continued fraction construction for the representation of a prime  $p \equiv 1 \pmod{4}$  as a sum of two squares is due to Legendre. Some other constructions are given in Chapter V of Davenport [17] and in Wagon [61]. A construction for the representation of any positive integer as a sum of four squares is given by Rousseau [46].

The modular group is the basic example of a *Fuchsian group*, i.e. a discrete subgroup of the group  $PSL_2(\mathbb{R})$  of all linear fractional transformations  $z \rightarrow (az + b)/(cz + d)$ , where  $a, b, c, d \in \mathbb{R}$  and  $ad - bc = 1$ . Fuchsian groups are studied from different points of view in the books of Katok [29], Beardon [7], Lehner [36], and Vinberg and Shvartsman [58].

The significance of Fuchsian groups stems in part from the uniformization theorem, which characterizes Riemann surfaces. A *Riemann surface* is a 1-dimensional complex manifold. Two Riemann surfaces are *conformally equivalent* if there is a bijective holomorphic map from one to the other. The *uniformization theorem*, first proved by Koebe and Poincaré independently in 1907, says that any Riemann surface is conformally equivalent to exactly one of the following:

- (i) the complex plane  $\mathbb{C}$ ,
- (ii) the Riemann sphere  $\mathbb{C} \cup \{\infty\}$ ,
- (iii) the cylinder  $\mathbb{C}/G$ , where  $G$  is the cyclic group generated by the translation  $z \rightarrow z + 1$ ,
- (iv) a torus  $\mathbb{C}/G$ , where  $G$  is the abelian group generated by the translations  $z \rightarrow z + 1$  and  $z \rightarrow z + \tau$  for some  $\tau \in \mathcal{H}$  (the upper half-plane),
- (v) a quotient space  $\mathcal{H}/G$ , where  $G$  is a Fuchsian group which acts *freely* on  $\mathcal{H}$ , i.e. if  $z \in \mathcal{H}$ ,  $g \in G$  and  $g \neq I$ , then  $g(z) \neq z$ .

(It should be noted that, since the modular group does not act freely on  $\mathcal{H}$ , the corresponding ‘Riemann surface’ is *ramified*.) For more information on the uniformization theorem, see Abikoff [1], Bers [9], Farkas and Kra [21], Jost [27], Beardon and Stephenson [8], and He and Schramm [24].

For the equivalence between quadratic fields and binary quadratic forms, see Zagier [63]. The class number  $h(d)$  of the quadratic field  $\mathbb{Q}(\sqrt{d})$  has been deeply investigated, originally by exploiting this equivalence. Dirichlet (1839) obtained an analytic formula for  $h(d)$  with the aid of his theorem on primes in an arithmetic progression (which will be proved in Chapter X). A clearly motivated proof of Dirichlet’s formula is given in Hasse [23], and there are some interesting observations on the formula in Stark [56].

It was conjectured by Gauss (1801), in the language of quadratic forms, that  $h(d) \rightarrow \infty$  as  $d \rightarrow -\infty$ . This was first proved by Heilbronn (1934). Siegel (1935) showed that actually

$$\log h(d)/\log |d| \rightarrow 1/2 \quad \text{as } d \rightarrow -\infty.$$

Generalizations of these results to arbitrary algebraic number fields are given in books on algebraic number theory, e.g. Narkiewicz [38].

Siegel (1943) has given a natural generalization of the modular group to higher dimensions. Instead of the upper half-plane  $\mathcal{H}$ , we consider the space  $\mathcal{H}_n$  of all complex  $n \times n$  matrices  $Z = X + iY$ , where  $X, Y$  are real symmetric matrices and  $Y$  is positive definite. If the real  $2n \times 2n$  matrix

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

is *symplectic*, i.e. if  $M^t J M = J$ , where

$$J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix},$$

then the linear fractional transformation  $Z \rightarrow (AZ + B)(CZ + D)^{-1}$ , maps  $\mathcal{H}_n$  onto itself. Siegel's modular group  $\Gamma_n$  is the group of all such transformations. The generalized upper half-plane  $\mathcal{H}_n$  is itself just a special case of the vast theory of symmetric Riemannian spaces initiated by E. Cartan (1926/7). See Siegel [54] and Helgason [25].

The development of non-Euclidean geometry is traced in Bonola [10]. (This edition also contains translations of works by Bolyai and Lobachevski.) The basic properties of Poincaré's model, here only stated, are proved in the books of Katok [29] and Beardon [7].

For the connection between continued fractions and geodesics, see Artin [5] and Sheingorn [53]. For the Markov spectrum see not only the books of Cassels [13] and Rockett and Szusz [45], but also Cusick and Flahive [16] and Baragar [6].

The theory of continued fractions for formal Laurent series is developed further in de Mathan [37]. The corresponding theory of Diophantine approximation is surveyed in Lasjaunias [35]. The polynomial Pell equation is discussed by Schmidt [48]. For formal Laurent series there is a multidimensional generalization which is quite different from those for real numbers; see Antoulas [4].

Roth's theorem and Schmidt's subspace theorem are proved in Schmidt [47]. See also Faltings and Wüstholz [20] and Evertse [19]. Nevanlinna's theory of the value distribution of meromorphic functions is treated in the recent book of Cherry and Ye [14]. For Vojta's work see, for example, [59] and [60]. It should be noted, though, that this area is still in a state of flux, besides using techniques beyond our scope. For an overview, see Lang [34].

Siegel's theorem on Diophantine equations with infinitely many solutions is proved with the aid of non-standard analysis by Robinson and Roquette [44]; the proof is reproduced in Stepanov [57]. The theorem is discussed from the standpoint of *Diophantine geometry* in Serre [50]. Any algebraic curve over  $\mathbb{Q}$  of genus zero which has a nonsingular rational point can be parametrized by rational functions *effectively*; see Poulakis [43].

It is worth noting that if  $F(x, y)$  is a polynomial with rational coefficients which is irreducible over  $\mathbb{Q}$ , but not over  $\mathbb{C}$ , then the curve  $F(x, y) = 0$  has at most finitely many rational points. For any rational point is a common root of at least two distinct complex-irreducible factors of  $F$  and any two such factors have at most finitely many common complex roots.

In conclusion we mention some further applications of continued fractions. A procedure, due to Vincent (1836), for separating the roots of a polynomial with integer coefficients has acquired some practical value with the advent of modern computers. See Alesina and Galuzzi [3].

Continued fractions play a role in the small divisor problems of classical mechanics. As an example, suppose the function  $f$  is holomorphic in some neighbourhood

of the origin and  $f(z) = \lambda z + O(z^2)$ , where  $\lambda = e^{2\pi i\theta}$  for some irrational  $\theta$ . It is readily shown that there exists a formal power series  $h$  which linearizes  $f$ , i.e.  $f(h(z)) = h(\lambda z)$ . Brjuno (1971) proved that this formal power series converges in a neighbourhood of the origin if  $\sum_{n \geq 0} (\log q_{n+1})/q_n < \infty$ , where  $q_n$  is the denominator of the  $n$ -th convergent of  $\theta$ . It was shown by Yoccoz (1995) that this condition is also necessary. In fact, if  $\sum_{n \geq 0} (\log q_{n+1})/q_n = \infty$ , the conclusion fails even for  $f(z) = \lambda z(1 - z)$ . See Yoccoz [62] and Pérez-Marco [40].

Our discussion of continued fractions has neglected their analytic theory. The outstanding work of Stieltjes (1894) on the *problem of moments*, which was extended by Hamburger (1920) and R. Nevanlinna (1922) from the half-line to the whole line, not only gave birth to the Stieltjes integral but also contributed to the development of functional analysis. For modern accounts, see Akhiezer [2], Landau [33] and Simon [55].

## 9 Selected References

- [1] W. Abikoff, The uniformization theorem, *Amer. Math. Monthly* **88** (1981), 574–592.
- [2] N.I. Akhiezer, *The classical moment problem*, Hafner, New York, 1965.
- [3] A. Alesina and M. Galuzzi, A new proof of Vincent's theorem, *Enseign. Math.* **44** (1998), 219–256.
- [4] A.C. Antoulas, On recursiveness and related topics in linear systems, *IEEE Trans. Automat. Control* **31** (1986), 1121–1135.
- [5] E. Artin, Ein mechanisches System mit quasiergodischen Bahnen, *Abh. Math. Sem. Univ. Hamburg* **3** (1924), 170–175. [*Collected Papers*, pp. 499–504, Addison-Wesley, Reading, Mass., 1965.]
- [6] A. Baragar, On the unicity conjecture for Markoff numbers, *Canad. Math. Bull.* **39** (1996), 3–9.
- [7] A.F. Beardon, *The geometry of discrete groups*, Springer-Verlag, New York, 1983.
- [8] A.F. Beardon and K. Stephenson, The uniformization theorem for circle packings, *Indiana Univ. Math. J.* **39** (1990), 1383–1425.
- [9] L. Bers, On Hilbert's 22nd problem, *Mathematical developments arising from Hilbert problems* (ed. F.E. Browder), pp. 559–609, Proc. Symp. Pure Math. **28**, Part 2, Amer. Math. Soc., Providence, R.I., 1976.
- [10] R. Bonola, *Non-Euclidean geometry*, English transl. by H.S. Carslaw, reprinted Dover, New York, 1955.
- [11] A.J. Brentjes, *Multi-dimensional continued fraction algorithms*, Mathematics Centre Tracts **145**, Amsterdam, 1981.
- [12] C. Brezinski, *History of continued fractions and Padé approximants*, Springer-Verlag, Berlin, 1991.
- [13] J.W.S. Cassels, *An introduction to Diophantine approximation*, Cambridge University Press, 1957.
- [14] W. Cherry and Z. Ye, *Nevanlinna's theory of value distribution*, Springer-Verlag, New York, 2000.
- [15] K.Y. Choong, D.E. Daykin and C.R. Rathbone, Rational approximations to  $\pi$ , *Math. Comp.* **25** (1971), 387–392.
- [16] T.W. Cusick and M.E. Flahive, *The Markoff and Lagrange spectra*, Mathematical Surveys and Monographs **30**, Amer. Math. Soc., Providence, R.I., 1989.
- [17] H. Davenport, *The higher arithmetic*, 7th ed., Cambridge University Press, 1999.

- [18] M. Davis, Y. Matijasevic and J. Robinson, Hilbert's tenth problem. Diophantine equations: positive aspects of a negative solution, *Mathematical developments arising from Hilbert problems* (ed. F.E. Browder), pp. 323–378, Proc. Symp. Pure Math. **28**, Part 2, Amer. Math. Soc., Providence, R.I., 1976.
- [19] J.H. Evertse, An improvement of the quantitative subspace theorem, *Compositio Math.* **101** (1996), 225–311.
- [20] G. Faltings and G. Wüstholz, Diophantine approximation on projective spaces, *Invent. Math.* **116** (1994), 109–138.
- [21] H.M. Farkas and I. Kra, *Riemann surfaces*, Springer-Verlag, New York, 1980.
- [22] H. Ferguson, A short proof of the existence of vector Euclidean algorithms, *Proc. Amer. Math. Soc.* **97** (1986), 8–10.
- [23] H. Hasse, *Vorlesungen über Zahlentheorie*, Zweite Auflage, Springer-Verlag, Berlin, 1964.
- [24] Z.-H. He and O. Schramm, On the convergence of circle packings to the Riemann map, *Invent. Math.* **125** (1996), 285–305.
- [25] S. Helgason, *Differential geometry, Lie groups, and symmetric spaces*, Academic Press, New York, 1978. [Corrected reprint, Amer. Math. Soc., Providence, R.I., 2001]
- [26] J.P. Jones and Y.V. Matijasevic, Proof of recursive unsolvability of Hilbert's tenth problem, *Amer. Math. Monthly* **98** (1991) 689–709.
- [27] J. Jost, *Compact Riemann surfaces*, transl. by R.R. Simha, Springer-Verlag, Berlin, 1997.
- [28] B. Just, Generalizing the continued fraction algorithm to arbitrary dimensions, *SIAM J. Comput.* **21** (1992), 909–926.
- [29] S. Katok, *Fuchsian groups*, University of Chicago Press, 1992.
- [30] J.F. Koksma, *Diophantische Approximationen*, Springer-Verlag, Berlin, 1936.
- [31] J.C. Lagarias, Geodesic multidimensional continued fractions, *Proc. London Math. Soc.* (3) **69** (1994), 464–488.
- [32] J.L. Lagrange, *Oeuvres*, t. VII, pp. 5–180, reprinted Olms Verlag, Hildesheim, 1973.
- [33] H.J. Landau, The classical moment problem: Hilbertian proofs, *J. Funct. Anal.* **38** (1980), 255–272.
- [34] S. Lang, *Number Theory III: Diophantine geometry*, Encyclopaedia of Mathematical Sciences Vol. 60, Springer-Verlag, Berlin, 1991.
- [35] A. Lasjaunias, A survey of Diophantine approximation in fields of power series, *Monatsh. Math.* **130** (2000), 211–229.
- [36] J. Lehner, *Discontinuous groups and automorphic functions*, Mathematical Surveys VIII, Amer. Math. Soc., Providence, R.I., 1964.
- [37] B. de Mathan, Approximations diophantiennes dans un corps local, *Bull. Soc. Math. France Suppl. Mém.* **21** (1970), Chapitre IV.
- [38] W. Narkiewicz, *Elementary and analytic theory of algebraic numbers*, 2nd ed., Springer-Verlag, Berlin, 1990.
- [39] W. Patz, *Tafel der regelmässigen Kettenbrüche und ihrer vollständigen Quotienten für die Quadratwurzeln aus den natürlichen Zahlen von 1–10000*, Akademie-Verlag, Berlin, 1955.
- [40] R. Pérez-Marco, Fixed points and circle maps, *Acta Math.* **179** (1997), 243–294.
- [41] O. Perron, *Die Lehre von den Kettenbrüchen*, Dritte Auflage, Teubner, Stuttgart, Band I, 1954; Band II, 1957. (Band II treats the analytic theory of continued fractions.)
- [42] E.V. Podsypanin, Length of the period of a quadratic irrational, *J. Soviet Math.* **18** (1982), 919–923.
- [43] D. Poulakis, Bounds for the minimal solution of genus zero Diophantine equations, *Acta Arith.* **86** (1998), 51–90.
- [44] A. Robinson and P. Roquette, On the finiteness theorem of Siegel and Mahler concerning Diophantine equations, *J. Number Theory* **7** (1975), 121–176.
- [45] A.M. Rockett and P. Szusz, *Continued fractions*, World Scientific, River Edge, N.J., 1992.
- [46] G. Rousseau, On a construction for the representation of a positive integer as the sum of four squares, *Enseign. Math.* (2) **33** (1987), 301–306.

- [47] W.M. Schmidt, *Diophantine approximation*, Lecture Notes in Mathematics **785**, Springer-Verlag, Berlin, 1980.
- [48] W.M. Schmidt, On continued fractions and diophantine approximation in power series fields, *Acta Arith.* **95** (2000), 139–166.
- [49] C.-O. Selenius, Rationale of the chakravala process of Jayadeva and Bhaskara II, *Historia Math.* **2** (1975), 167–184.
- [50] J.-P. Serre, *Lectures on the Mordell–Weil theorem*, English transl. by M. Brown from notes by M. Waldschmidt, Vieweg & Sohn, Braunschweig, 1989.
- [51] J.A. Serret, Developpements sur une classe d'équations, *J. Math. Pures Appl.* **15** (1850), 152–168.
- [52] J. Shallit, Real numbers with bounded partial quotients, *Enseign. Math.* **38** (1992), 151–187.
- [53] M. Sheingorn, Continued fractions and congruence subgroup geodesics, *Number theory with an emphasis on the Markoff spectrum* (ed. A.D. Pollington and W. Moran), pp. 239–254, Lecture Notes in Pure and Applied Mathematics **147**, Dekker, New York, 1993.
- [54] C.L. Siegel, Symplectic geometry, *Amer. J. Math.* **65** (1943), 1–86. [*Gesammelte Abhandlungen, Band II*, pp. 274–359, Springer-Verlag, Berlin, 1966.]
- [55] B. Simon, The classical moment problem as a self-adjoint finite difference operator, *Adv. in Math.* **137** (1998), 82–203.
- [56] H.M. Stark, Dirichlet's class-number formula revisited, *A tribute to Emil Grosswald: Number theory and related analysis* (ed. M. Knopp and M. Sheingorn), pp. 571–577, Contemporary Mathematics **143**, Amer. Math. Soc., Providence, R.I., 1993.
- [57] S.A. Stepanov, *Arithmetic of algebraic curves*, English transl. by I. Aleksanova, Consultants Bureau, New York, 1994.
- [58] E.B. Vinberg and O.V. Shvartsman, *Discrete groups of motions of spaces of constant curvature*, Geometry II, pp. 139–248, Encyclopaedia of Mathematical Sciences Vol. 29, Springer-Verlag, Berlin, 1993.
- [59] P. Vojta, *Diophantine approximations and value distribution theory*, Lecture Notes in Mathematics **1239**, Springer-Verlag, Berlin, 1987.
- [60] P. Vojta, A generalization of theorems of Faltings and Thue–Siegel–Roth–Wirsing, *J. Amer. Math. Soc.* **5** (1992), 763–804.
- [61] S. Wagon, The Euclidean algorithm strikes again, *Amer. Math. Monthly* **97** (1990), 125–129.
- [62] J.-C. Yoccoz, Théorème de Siegel, nombres de Bruno et polynômes quadratiques, *Astérisque* **231** (1995), 3–88.
- [63] D.B. Zagier, *Zetafunktionen und quadratische Körper*, Springer-Verlag, Berlin, 1981.

## Additional References

- M. Laczkovich, On Lambert's proof of the irrationality of  $\pi$ , *Amer. Math. Monthly* **104** (1997), 439–443.
- Anitha Srinivasan, A really simple proof of the Markoff conjecture for prime powers, *Preprint*.

# Hadamard’s Determinant Problem

It was shown by Hadamard (1893) that, if all elements of an  $n \times n$  matrix of complex numbers have absolute value at most  $\mu$ , then the determinant of the matrix has absolute value at most  $\mu^n n^{n/2}$ . For each positive integer  $n$  there exist complex  $n \times n$  matrices for which this upper bound is attained. For example, the upper bound is attained for  $\mu = 1$  by the matrix  $(\omega^{jk})(1 \leq j, k \leq n)$ , where  $\omega$  is a primitive  $n$ -th root of unity. This matrix is real for  $n = 1, 2$ . However, Hadamard also showed that if the upper bound is attained for a real  $n \times n$  matrix, where  $n > 2$ , then  $n$  is divisible by 4.

Without loss of generality one may suppose  $\mu = 1$ . A real  $n \times n$  matrix for which the upper bound  $n^{n/2}$  is attained in this case is today called a *Hadamard matrix*. It is still an open question whether an  $n \times n$  Hadamard matrix exists for every positive integer  $n$  divisible by 4.

Hadamard’s inequality played an important role in the theory of linear integral equations created by Fredholm (1900), and partly for this reason many proofs and generalizations were soon given. Fredholm’s approach to linear integral equations has been superseded, but Hadamard’s inequality has found connections with several other branches of mathematics, such as number theory, combinatorics and group theory. Hadamard matrices have been used to enhance the precision of spectrometers, to design agricultural experiments and to correct errors in messages transmitted by spacecraft.

The moral is that a good mathematical problem will in time find applications. Although the case where  $n$  is divisible by 4 has a richer theory, we will also treat other cases of Hadamard’s determinant problem, since progress with them might lead to progress also for Hadamard matrices.

## 1 What is a Determinant?

The system of two simultaneous linear equations

$$a_{11}\zeta_1 + a_{12}\zeta_2 = \beta_1$$

$$a_{21}\zeta_1 + a_{22}\zeta_2 = \beta_2$$

has, if  $\delta_2 = \alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}$  is nonzero, the unique solution

$$\zeta_1 = (\beta_1\alpha_{22} - \beta_2\alpha_{12})/\delta_2, \quad \zeta_2 = -(\beta_1\alpha_{21} - \beta_2\alpha_{11})/\delta_2.$$

If  $\delta_2 = 0$ , then either there is no solution or there is more than one solution.

Similarly the system of three simultaneous linear equations

$$\begin{aligned}\alpha_{11}\zeta_1 + \alpha_{12}\zeta_2 + \alpha_{13}\zeta_3 &= \beta_1 \\ \alpha_{21}\zeta_1 + \alpha_{22}\zeta_2 + \alpha_{23}\zeta_3 &= \beta_2 \\ \alpha_{31}\zeta_1 + \alpha_{32}\zeta_2 + \alpha_{33}\zeta_3 &= \beta_3\end{aligned}$$

has a unique solution if and only if  $\delta_3 \neq 0$ , where

$$\begin{aligned}\delta_3 &= \alpha_{11}\alpha_{22}\alpha_{33} + \alpha_{12}\alpha_{23}\alpha_{31} + \alpha_{13}\alpha_{21}\alpha_{32} \\ &\quad - \alpha_{11}\alpha_{23}\alpha_{32} - \alpha_{12}\alpha_{21}\alpha_{33} - \alpha_{13}\alpha_{22}\alpha_{31}.\end{aligned}$$

These considerations may be extended to any finite number of simultaneous linear equations. The system

$$\begin{aligned}\alpha_{11}\zeta_1 + \alpha_{12}\zeta_2 + \cdots + \alpha_{1n}\zeta_n &= \beta_1 \\ \alpha_{21}\zeta_1 + \alpha_{22}\zeta_2 + \cdots + \alpha_{2n}\zeta_n &= \beta_2 \\ &\vdots \\ \alpha_{n1}\zeta_1 + \alpha_{n2}\zeta_2 + \cdots + \alpha_{nn}\zeta_n &= \beta_n\end{aligned}$$

has a unique solution if and only if  $\delta_n \neq 0$ , where

$$\delta_n = \sum \pm \alpha_{1k_1}\alpha_{2k_2} \cdots \alpha_{nk_n},$$

the sum being taken over all  $n!$  permutations  $k_1, k_2, \dots, k_n$  of  $1, 2, \dots, n$  and the sign chosen being  $+$  or  $-$  according as the permutation is even or odd, as defined in Chapter I, §7.

It has been tacitly assumed that the given quantities  $\alpha_{jk}, \beta_j (j, k = 1, \dots, n)$  are real numbers, in which case the solution  $\zeta_k (k = 1, \dots, n)$  also consists of real numbers. However, everything that has been said remains valid if the given quantities are elements of an arbitrary field  $F$ , in which case the solution also consists of elements of  $F$ . Since  $\delta_n$  is an element of  $F$  which is uniquely determined by the matrix

$$A = \begin{bmatrix} \alpha_{11} & \cdots & \alpha_{1n} \\ \vdots & \ddots & \vdots \\ \alpha_{n1} & \cdots & \alpha_{nn} \end{bmatrix},$$

it will be called the *determinant* of the matrix  $A$  and denoted by  $\det A$ .

Determinants appear in the work of the Japanese mathematician Seki (1683) and in a letter of Leibniz (1693) to l'Hospital, but neither had any influence on later developments. The rule which expresses the solution of a system of linear equations by quotients of determinants was stated by Cramer (1750), but the study of determinants for their own sake began with Vandermonde (1771). The word 'determinant' was first used in the present sense by Cauchy (1812), who gave a systematic account of their

theory. The diffusion of this theory throughout the mathematical world owes much to the clear exposition of Jacobi (1841).

For the practical solution of linear equations Cramer's rule is certainly inferior to the age-old method of elimination of variables. Even many of the theoretical uses to which determinants were once put have been replaced by simpler arguments from linear algebra, to the extent that some have advocated banning determinants from the curriculum. However, determinants have a geometrical interpretation which makes their survival desirable.

Let  $M_n(\mathbb{R})$  denote the set of all  $n \times n$  matrices with entries from the real field  $\mathbb{R}$ . If  $A \in M_n(\mathbb{R})$ , then the linear map  $x \rightarrow Ax$  of  $\mathbb{R}^n$  into itself multiplies the volume of any parallelotope by a fixed factor  $\mu(A) \geq 0$ . Evidently

- (i)''  $\mu(AB) = \mu(A)\mu(B)$  for all  $A, B \in M_n(\mathbb{R})$ ,
- (ii)''  $\mu(D) = |\alpha|$  for any diagonal matrix  $D = \text{diag}[1, \dots, 1, \alpha] \in M_n(\mathbb{R})$ .

(A matrix  $A = (\alpha_{jk})$  is denoted by  $\text{diag}[\alpha_{11}, \alpha_{22}, \dots, \alpha_{nn}]$  if  $\alpha_{jk} = 0$  whenever  $j \neq k$  and is then said to be *diagonal*.) It may be shown (e.g., by representing  $A$  as a product of elementary matrices in the manner described below) that  $\mu(A) = |\det A|$ . The sign of the determinant also has a geometrical interpretation:  $\det A \geq 0$  according as the linear map  $x \rightarrow Ax$  preserves or reverses orientation.

Now let  $F$  be an arbitrary field and let  $M_n = M_n(F)$  denote the set of all  $n \times n$  matrices with entries from  $F$ . We intend to show that determinants, as defined above, have the properties:

- (i)'  $\det(AB) = \det A \cdot \det B$  for all  $A, B \in M_n$ ,
- (ii)'  $\det D = \alpha$  for any diagonal matrix  $D = \text{diag}[1, \dots, 1, \alpha] \in M_n$ ,

and, moreover, that these two properties actually characterize determinants. To avoid notational complexity, we consider first the case  $n = 2$ .

Let  $\mathcal{E}$  denote the set of all matrices  $A \in M_2$  which are products of finitely many matrices of the form  $U_\lambda, V_\mu$ , where

$$U_\lambda = \begin{bmatrix} 1 & \lambda \\ 0 & 1 \end{bmatrix}, \quad V_\mu = \begin{bmatrix} 1 & 0 \\ \mu & 1 \end{bmatrix},$$

and  $\lambda, \mu \in F$ . The set  $\mathcal{E}$  is a group under matrix multiplication, since multiplication is associative,  $I \in \mathcal{E}$ ,  $\mathcal{E}$  is obviously closed under multiplication and  $U_\lambda, V_\mu$  have inverses  $U_{-\lambda}, V_{-\mu}$  respectively.

We are going to show that, if  $A \in M_2$  and  $A \neq O$ , then there exist  $S, T \in \mathcal{E}$  and  $\delta \in F$  such that  $SAT = \text{diag}[1, \delta]$ .

For any  $\rho \neq 0$ , put

$$W = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad R_\rho = \begin{bmatrix} \rho^{-1} & 0 \\ 0 & \rho \end{bmatrix}.$$

Then  $W = U_{-1}V_1U_{-1} \in \mathcal{E}$  and also  $R_\rho \in \mathcal{E}$  since, if  $\sigma = 1 - \rho$ ,  $\rho' = \rho^{-1}$  and  $\tau = \rho^2 - \rho$ , then

$$R_\rho = V_{-1}U_\sigma V_{\rho'}U_\tau.$$

Let

$$A = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix},$$

where at least one of  $\alpha, \beta, \gamma, \delta$  is nonzero. By multiplying  $A$  on the left, or on the right, or both by  $W$  we may suppose that  $\alpha \neq 0$ . Now, by multiplying  $A$  on the right or left by  $R_\alpha$ , we may suppose that  $\alpha = 1$ . Next, by multiplying  $A$  on the right by  $U_{-\beta}$ , we may further suppose that  $\beta = 0$ . Finally, by multiplying  $A$  on the left by  $V_{-\gamma}$ , we may also suppose that  $\gamma = 0$ .

The preceding argument is valid even if  $F$  is a division ring. In what follows we will use the commutativity of multiplication in  $F$ .

We are now going to show that if  $d : \mathcal{E} \rightarrow F$  is a map such that  $d(ST) = d(S)d(T)$  for all  $S, T \in \mathcal{E}$ , then either  $d(S) = 0$  for every  $S \in \mathcal{E}$  or  $d(S) = 1$  for every  $S \in \mathcal{E}$ .

If  $d(T) = 0$  for some  $T \in \mathcal{E}$ , then  $d(I) = d(T)d(T^{-1}) = 0$  and  $d(S) = d(I)d(S) = 0$  for every  $S \in \mathcal{E}$ . Thus we now suppose  $d(S) \neq 0$  for every  $S \in \mathcal{E}$ . Then, in the same way,  $d(I) = 1$  and  $d(S^{-1}) = d(S)^{-1}$  for every  $S \in \mathcal{E}$ .

It is easily verified that

$$\begin{aligned} U_\lambda U_\mu &= U_{\lambda+\mu}, & V_\lambda V_\mu &= V_{\lambda+\mu}, \\ W^{-1} &= -W, & W^{-1} V_\mu W &= U_{-\mu}. \end{aligned}$$

It follows that

$$d(V_\mu) = d(U_{-\mu}) = d(U_\mu)^{-1}.$$

Also, for any  $\rho \neq 0$ ,

$$R_\rho^{-1} U_\lambda R_\rho = U_{\lambda\rho^2}.$$

Hence  $d(U_{\lambda\rho^2}) = d(U_\lambda)$  and  $d(U_{\lambda(\rho^2-1)}) = 1$ .

If the field  $F$  contains more than three elements, then  $\rho^2 - 1 \neq 0$  for some nonzero  $\rho \in F$ . Since  $\lambda(\rho^2 - 1)$  runs through the nonzero elements of  $F$  at the same time as  $\lambda$ , it follows that  $d(U_\lambda) = 1$  for every  $\lambda \in F$ . Hence also  $d(V_\mu) = 1$  for every  $\mu \in F$  and  $d(S) = 1$  for all  $S \in \mathcal{E}$ .

If  $F$  contains 2 elements, then  $d(S) = 1$  for every  $S \in \mathcal{E}$  is the only possibility. If  $F$  contains 3 elements, then  $d(S) = \pm 1$  for every  $S \in \mathcal{E}$ . Hence  $d(S^{-1}) = d(S)$  and  $d(S^2) = 1$ . Since  $U_2 = U_1^2$  and  $U_1 = U_2^{-1}$ , this implies  $d(U_\lambda) = 1$  for every  $\lambda \in F$ , and the rest follows as before.

The preceding discussion is easily extended to higher dimensions. Put

$$U_{ij}(\lambda) = I_n + \lambda E_{ij},$$

for any  $i, j \in \{1, \dots, n\}$  with  $i \neq j$ , where  $E_{ij}$  is the  $n \times n$  matrix with all entries 0 except the  $(i, j)$ -th, which is 1, and let  $SL_n(F)$  denote the set of all  $A \in M_n$  which are products of finitely many matrices  $U_{ij}(\lambda)$ . Then  $SL_n(F)$  is a group under matrix multiplication.

If  $A \in M_n$  and  $A \neq O$ , then there exist  $S, T \in SL_n(F)$  and a positive integer  $r \leq n$  such that

$$SAT = \text{diag}[1_{r-1}, \delta, 0_{n-r}]$$

for some nonzero  $\delta \in F$ . The matrix  $A$  is *singular* if  $r < n$  and *nonsingular* if  $r = n$ . Hence  $A = (\alpha_{jk})$  is nonsingular if and only if its transpose  $A^t = (\alpha_{kj})$  is nonsingular. In the nonsingular case we need multiply  $A$  on only one side by a matrix from  $SL_n(F)$  to bring it to the form

$$D_\delta = \text{diag}[1_{n-1}, \delta].$$

For if  $SAT = D_\delta$ , then  $SA = D_\delta T^{-1}$  and this implies  $SA = S'D_\delta$  for some  $S' \in SL_n(F)$ , since

$$\begin{aligned} D_\delta U_{ij}(\lambda) &= U_{ij}(\lambda\delta^{-1})D_\delta & \text{if } i < j = n, \\ D_\delta U_{ij}(\lambda) &= U_{ij}(\delta\lambda)D_\delta & \text{if } j < i = n, \\ D_\delta U_{ij}(\lambda) &= U_{ij}(\lambda)D_\delta & \text{if } i, j \neq n \text{ and } i \neq j. \end{aligned}$$

In the same way as for  $n = 2$  it may be shown that, if  $d : SL_n(F) \rightarrow F$  is a map such that  $d(ST) = d(S)d(T)$  for all  $S, T \in SL_n(F)$ , then either  $d(S) = 0$  for every  $S$  or  $d(S) = 1$  for every  $S$ .

**Theorem 1** *There exists a unique map  $d : M_n \rightarrow F$  such that*

- (i)'  $d(AB) = d(A)d(B)$  for all  $A, B \in M_n$ ,
- (ii)' for any  $\alpha \in F$ , if  $D_\alpha = \text{diag}[1_{n-1}, \alpha]$ , then  $d(D_\alpha) = \alpha$ .

*Proof* We consider first uniqueness. Since  $d(I) = d(D_1) = 1$ , we must have  $d(S) = 1$  for every  $S \in SL_n(F)$ , by what we have just said. Also, if

$$H = \text{diag}[\eta_1, \dots, \eta_{n-1}, 0],$$

then  $d(H) = 0$ , since  $H = D_0H$ . In particular,  $d(O) = 0$ . If  $A \in M_n$  and  $A \neq O$ , there exist  $S, T \in SL_n(F)$  such that

$$SAT = \text{diag}[1_{r-1}, \delta, 0_{n-r}],$$

where  $1 \leq r \leq n$  and  $\delta \neq 0$ . It follows that  $d(A) = 0$  if  $r < n$ , i.e. if  $A$  is singular. On the other hand if  $r = n$ , i.e. if  $A$  is nonsingular, then  $SAT = D_\delta$  and hence  $d(A) = \delta$ . This proves uniqueness.

We consider next existence. For any  $A = (\alpha_{jk}) \in M_n$ , define

$$\det A = \sum_{\sigma \in \mathcal{S}_n} (\text{sgn } \sigma) \alpha_{1\sigma 1} \alpha_{2\sigma 2} \cdots \alpha_{n\sigma n},$$

where  $\sigma$  is a permutation of  $1, 2, \dots, n$ ,  $\text{sgn } \sigma = 1$  or  $-1$  according as the permutation  $\sigma$  is even or odd, and the summation is over the symmetric group  $\mathcal{S}_n$  of all permutations. Several consequences of this definition will now be derived.

(i) if every entry in some row of  $A$  is 0, then  $\det A = 0$ .

*Proof* Every summand vanishes in the expression for  $\det A$ . □

(ii) if the matrix  $B$  is obtained from the matrix  $A$  by multiplying all entries in one row by  $\lambda$ , then  $\det B = \lambda \det A$ .

*Proof* This is also clear, since in the expression for  $\det A$  each summand contains exactly one factor from any given row.  $\square$

(iii) if two rows of  $A$  are the same, then  $\det A = 0$ .

*Proof* Suppose for definiteness that the first and second rows are the same, and let  $\tau$  be the permutation which interchanges 1 and 2 and leaves fixed every  $k > 2$ . Then  $\tau$  is odd and we can write

$$\det A = \sum_{\sigma \in \mathcal{A}_n} \alpha_{1\sigma 1} \alpha_{2\sigma 2} \cdots \alpha_{n\sigma n} - \sum_{\sigma \in \mathcal{A}_n} \alpha_{1\sigma \tau 1} \alpha_{2\sigma \tau 2} \cdots \alpha_{n\sigma \tau n},$$

where  $\mathcal{A}_n$  is the alternating group of all even permutations. In the second sum

$$\alpha_{1\sigma \tau 1} \alpha_{2\sigma \tau 2} \cdots \alpha_{n\sigma \tau n} = \alpha_{1\sigma 2} \alpha_{2\sigma 1} \alpha_{3\sigma 3} \cdots \alpha_{n\sigma n} = \alpha_{2\sigma 2} \alpha_{1\sigma 1} \alpha_{3\sigma 3} \cdots \alpha_{n\sigma n},$$

because the first and second rows are the same. Hence the two sums cancel.  $\square$

(iv) if the matrix  $B$  is obtained from the matrix  $A$  by adding a scalar multiple of one row to a different row, then  $\det B = \det A$ .

*Proof* Suppose for definiteness that  $B$  is obtained from  $A$  by adding  $\lambda$  times the second row to the first. Then

$$\det B = \sum_{\sigma \in \mathcal{S}_n} (\operatorname{sgn} \sigma) \alpha_{1\sigma 1} \alpha_{2\sigma 2} \cdots \alpha_{n\sigma n} + \lambda \sum_{\sigma \in \mathcal{S}_n} (\operatorname{sgn} \sigma) \alpha_{2\sigma 1} \alpha_{2\sigma 2} \cdots \alpha_{n\sigma n}.$$

The first sum is  $\det A$  and the second sum is 0, by (iii), since it is the determinant of the matrix obtained from  $A$  by replacing the first row by the second.  $\square$

(v) if  $A$  is singular, then  $\det A = 0$ .

*Proof* If  $A$  is singular, then some row of  $A$  is a linear combination of the remaining rows. Thus by subtracting from this row scalar multiples of the remaining rows we can replace it by a row of 0's. For the new matrix  $B$  we have  $\det B = 0$ , by (i). On the other hand,  $\det B = \det A$ , by (iv).  $\square$

(vi) if  $A = \operatorname{diag}[\delta_1, \dots, \delta_n]$ , then  $\det A = \delta_1 \cdots \delta_n$ . In particular,  $\det D_\alpha = \alpha$ .

*Proof* In the expression for  $\det A$  the only possible nonzero summand is that for which  $\sigma$  is the identity permutation, and the identity permutation is even.  $\square$

(vii)  $\det(AB) = \det A \cdot \det B$  for all  $A, B \in M_n$ .

*Proof* If  $A$  is singular, then  $AB$  is also and so, by (v),  $\det(AB) = 0 = \det A \cdot \det B$ . Thus we now suppose that  $A$  is nonsingular. Then there exists  $S \in SL_n(F)$  such that  $SA = D_\delta$  for some nonzero  $\delta \in F$ . Since, by the definition of  $SL_n(F)$ , left multiplication by  $S$  corresponds to a finite number of operations of the type considered in (iv) we have

$$\det A = \det(SA) = \det D_\delta$$

and

$$\det(AB) = \det(SAB) = \det(D_\delta B).$$

But  $\det D_\delta = \delta$ , by (vi), and  $\det(D_\delta B) = \delta \det B$ , by (ii). Therefore  $\det(AB) = \det A \cdot \det B$ .

This completes the proof of existence.  $\square$

**Corollary 2** *If  $A \in M_n$  and if  $A^t$  is the transpose of  $A$ , then  $\det A^t = \det A$ .*

*Proof* The map  $d : M_n \rightarrow F$  defined by  $d(A) = \det A^t$  also has the properties (i)', (ii)'.  $\square$

The proof of Theorem 1 shows further that  $SL_n(F)$  is the special linear group, consisting of all  $A \in M_n$  with  $\det A = 1$ .

We do not propose to establish here all the properties of determinants which we may later require. However, we note that if

$$A = \begin{bmatrix} B & 0 \\ C & D \end{bmatrix}$$

is a partitioned matrix, where  $B$  and  $D$  are square matrices of smaller size, then

$$\det A = \det B \cdot \det D.$$

It follows that if  $A = (a_{jk})$  is *lower triangular* (i.e.  $a_{jk} = 0$  for all  $j, k$  with  $j < k$ ) or *upper triangular* (i.e.  $a_{jk} = 0$  for all  $j, k$  with  $j > k$ ), then

$$\det A = a_{11}a_{22} \cdots a_{nn}.$$

## 2 Hadamard Matrices

We begin by obtaining an upper bound for  $\det(A^t A)$ , where  $A$  is an  $n \times m$  real matrix. If  $m = n$ , then  $\det(A^t A) = (\det A)^2$  and bounding  $\det(A^t A)$  is the same as Hadamard's problem of bounding  $|\det A|$ . However, as we will see in §3, the problem is of interest also for  $m < n$ .

In the statement of the following result we denote by  $\|v\|$  the Euclidean norm of a vector  $v = (a_1, \dots, a_n) \in \mathbb{R}^n$ . Thus  $\|v\| \geq 0$  and  $\|v\|^2 = a_1^2 + \cdots + a_n^2$ . The geometrical interpretation of the result is that a parallelotope with given side lengths has maximum volume when the sides are orthogonal.

**Proposition 3** *Let  $A$  be an  $n \times m$  real matrix with linearly independent columns  $v_1, \dots, v_m$ . Then*

$$\det(A^t A) \leq \prod_{k=1}^m \|v_k\|^2,$$

*with equality if and only if  $A^t A$  is a diagonal matrix.*

*Proof* We are going to construct inductively mutually orthogonal vectors  $w_1, \dots, w_m$  such that  $w_k$  is a linear combination of  $v_1, \dots, v_k$  in which the coefficient of  $v_k$  is 1 ( $1 \leq k \leq m$ ). Take  $w_1 = v_1$  and suppose  $w_1, \dots, w_{k-1}$  have been determined. If we take

$$w_k = v_k - \alpha_1 w_1 - \dots - \alpha_{k-1} w_{k-1},$$

where  $\alpha_j = \langle v_k, w_j \rangle$ , then  $\langle w_k, w_j \rangle = 0$  ( $1 \leq j < k$ ). Moreover,  $w_k \neq 0$ , since  $v_1, \dots, v_k$  are linearly independent. (This is the same process as in §10 of Chapter I, but without the normalization.)

If  $B$  is the matrix with columns  $w_1, \dots, w_m$  then, by construction,

$$B^t B = \text{diag}[\delta_1, \dots, \delta_m]$$

is a diagonal matrix with diagonal entries  $\delta_k = \|w_k\|^2$  and  $AT = B$  for some upper triangular matrix  $T$  with 1's in the main diagonal. Since  $\det T = 1$ , we have

$$\det(A^t A) = \det(B^t B) = \prod_{k=1}^m \|w_k\|^2.$$

But

$$\|v_k\|^2 = \|w_k\|^2 + |\alpha_1|^2 \|w_1\|^2 + \dots + |\alpha_{k-1}|^2 \|w_{k-1}\|^2$$

and hence  $\|w_k\|^2 \leq \|v_k\|^2$ , with equality only if  $w_k = v_k$ . The result follows.  $\square$

**Corollary 4** Let  $A = (a_{jk})$  be an  $n \times m$  real matrix such that  $|a_{jk}| \leq 1$  for all  $j, k$ . Then

$$\det(A^t A) \leq n^m,$$

with equality if and only if  $a_{jk} = \pm 1$  for all  $j, k$  and  $A^t A = nI_m$ .

*Proof* We may assume that the columns of  $A$  are linearly independent, since otherwise  $\det(A^t A) = 0$ . If  $v_k$  is the  $k$ -th column of  $A$ , then  $\|v_k\|^2 \leq n$ , with equality if and only if  $|a_{jk}| = 1$  for  $1 \leq j \leq n$ . The result now follows from Proposition 3.  $\square$

An  $n \times m$  matrix  $A = (a_{jk})$  will be said to be an *H-matrix* if  $a_{jk} = \pm 1$  for all  $j, k$  and  $A^t A = nI_m$ . If, in addition,  $m = n$  then  $A$  will be said to be a *Hadamard matrix* of order  $n$ .

If  $A$  is an  $n \times m$  *H-matrix*, then  $m \leq n$ . Furthermore, if  $A$  is a Hadamard matrix of order  $n$  then, for any  $m < n$ , the submatrix formed by the first  $m$  columns of  $A$  is an *H-matrix*. (This distinction between *H-matrices* and Hadamard matrices is convenient, but not standard. It is an unproven conjecture that any *H-matrix* can be completed to a Hadamard matrix.)

The transpose  $A^t$  of a Hadamard matrix  $A$  is again a Hadamard matrix, since  $A^t = nA^{-1}$  commutes with  $A$ . The  $1 \times 1$  unit matrix is a Hadamard matrix, and so is the  $2 \times 2$  matrix

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

There is one rather simple procedure for constructing  $H$ -matrices. If  $A = (\alpha_{jk})$  is an  $n \times m$  matrix and  $B = (\beta_{i\ell})$  a  $q \times p$  matrix, then the  $nq \times mp$  matrix

$$\begin{bmatrix} \alpha_{11}B & \alpha_{12}B & \cdots & \alpha_{1m}B \\ \alpha_{21}B & \alpha_{22}B & \cdots & \alpha_{2m}B \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_{n1}B & \alpha_{n2}B & \cdots & \alpha_{nm}B \end{bmatrix},$$

with entries  $\alpha_{jk}\beta_{i\ell}$ , is called the *Kronecker product* of  $A$  and  $B$  and is denoted by  $A \otimes B$ . It is easily verified that

$$(A \otimes B)(C \otimes D) = AC \otimes BD$$

and

$$(A \otimes B)^t = A^t \otimes B^t.$$

It follows directly from these rules of calculation that if  $A_1$  is an  $n_1 \times m_1$   $H$ -matrix and  $A_2$  an  $n_2 \times m_2$   $H$ -matrix, then  $A_1 \otimes A_2$  is an  $n_1n_2 \times m_1m_2$   $H$ -matrix. Consequently, since there exist Hadamard matrices of orders 1 and 2, there also exist Hadamard matrices of order any power of 2. This was already known to Sylvester (1867).

**Proposition 5** *Let  $A = (\alpha_{jk})$  be an  $n \times m$   $H$ -matrix. If  $n > 1$ , then  $n$  is even and any two distinct columns of  $A$  have the same entries in exactly  $n/2$  rows. If  $n > 2$ , then  $n$  is divisible by 4 and any three distinct columns of  $A$  have the same entries in exactly  $n/4$  rows.*

*Proof* If  $j \neq k$ , then

$$\alpha_{1j}\alpha_{1k} + \cdots + \alpha_{nj}\alpha_{nk} = 0.$$

Since  $\alpha_{ij}\alpha_{ik} = 1$  if the  $j$ -th and  $k$ -th columns have the same entry in the  $i$ -th row and  $= -1$  otherwise, the number of rows in which the  $j$ -th and  $k$ -th columns have the same entry is  $n/2$ .

If  $j, k, \ell$  are all different, then

$$\sum_{i=1}^n (\alpha_{ij} + \alpha_{ik})(\alpha_{ij} + \alpha_{i\ell}) = \sum_{i=1}^n \alpha_{ij}^2 = n.$$

But  $(\alpha_{ij} + \alpha_{ik})(\alpha_{ij} + \alpha_{i\ell}) = 4$  if the  $j$ -th,  $k$ -th and  $\ell$ -th columns all have the same entry in the  $i$ -th row and  $= 0$  otherwise. Hence the number of rows in which the  $j$ -th,  $k$ -th and  $\ell$ -th columns all have the same entry is exactly  $n/4$ .  $\square$

Thus the order  $n$  of a Hadamard matrix must be divisible by 4 if  $n > 2$ . It is unknown if a Hadamard matrix of order  $n$  exists for every  $n$  divisible by 4. However, it is known for  $n \leq 424$  and for several infinite families of  $n$ . We restrict attention here to the family of Hadamard matrices constructed by Paley (1933).

The following lemma may be immediately verified by matrix multiplication.

**Lemma 6** Let  $C$  be an  $n \times n$  matrix, with 0's on the main diagonal and all other entries 1 or  $-1$ , such that

$$C^t C = (n-1)I_n.$$

If  $C$  is skew-symmetric (i.e.  $C^t = -C$ ), then  $C + I$  is a Hadamard matrix of order  $n$ , whereas if  $C$  is symmetric (i.e.  $C^t = C$ ), then

$$\begin{bmatrix} C + I & C - I \\ C - I & -C - I \end{bmatrix}$$

is a Hadamard matrix of order  $2n$ .

**Proposition 7** If  $q$  is a power of an odd prime, there exists a  $(q+1) \times (q+1)$  matrix  $C$  with 0's on the main diagonal and all other entries 1 or  $-1$ , such that

- (i)  $C^t C = qI_{q+1}$ ,
- (ii)  $C$  is skew-symmetric if  $q \equiv 3 \pmod{4}$  and symmetric if  $q \equiv 1 \pmod{4}$ .

*Proof* Let  $F$  be a finite field containing  $q$  elements. Since  $q$  is odd, not all elements of  $F$  are squares. For any  $a \in F$ , put

$$\chi(a) = \begin{cases} 0 & \text{if } a = 0, \\ 1 & \text{if } a \neq 0 \text{ and } a = c^2 \text{ for some } c \in F, \\ -1 & \text{if } a \text{ is not a square.} \end{cases}$$

If  $q = p$  is a prime, then  $F$  is the field of integers modulo  $p$  and  $\chi(a) = (a/p)$  is the Legendre symbol studied in Chapter III. The following argument may be restricted to this case, if desired.

Since the multiplicative group of  $F$  is cyclic, we have

$$\chi(ab) = \chi(a)\chi(b) \quad \text{for all } a, b \in F.$$

Since the number of nonzero elements which are squares is equal to the number which are non-squares, we also have

$$\sum_{a \in F} \chi(a) = 0.$$

It follows that, for any  $c \neq 0$ ,

$$\sum_{b \in F} \chi(b)\chi(b+c) = \sum_{b \neq 0} \chi(b)^2 \chi(1+cb^{-1}) = \sum_{x \neq 1} \chi(x) = -1.$$

Let  $0 = a_0, a_1, \dots, a_{q-1}$  be an enumeration of the elements of  $F$  and define a  $q \times q$  matrix  $Q = (q_{jk})$  by

$$q_{jk} = \chi(a_j - a_k) \quad (0 \leq j, k < q).$$

Thus  $Q$  has 0's on the main diagonal and  $\pm 1$ 's elsewhere. Also, by what has been said in the previous paragraph, if  $J_m$  denotes the  $m \times m$  matrix with all entries 1, then

$$QJ_q = 0, \quad Q^t Q = qI_q - J_q.$$

Furthermore, since  $\chi(-1) = (-1)^{(q-1)/2}$ ,  $Q$  is symmetric if  $q \equiv 1 \pmod{4}$  and skew-symmetric if  $q \equiv 3 \pmod{4}$ . If  $e_m$  denotes the  $1 \times m$  matrix with all entries 1, it follows that the matrix

$$C = \begin{bmatrix} 0 & e_q \\ \pm e_q^t & Q \end{bmatrix},$$

where the  $\pm$  sign is chosen according as  $q \equiv \pm 1 \pmod{4}$ , satisfies the various requirements.  $\square$

By combining Lemma 6 with Proposition 7 we obtain Paley's result that, for any odd prime power  $q$ , there exists a Hadamard matrix of order  $q + 1$  if  $q \equiv 3 \pmod{4}$  and of order  $2(q + 1)$  if  $q \equiv 1 \pmod{4}$ . Together with the Kronecker product construction, this establishes the existence of Hadamard matrices for all orders  $n \equiv 0 \pmod{4}$  with  $n \leq 100$ , except  $n = 92$ .

A Hadamard matrix of order 92 was found by Baumert, Golomb and Hall (1962), using a computer search and the following method proposed by Williamson (1944). Let  $A, B, C, D$  be  $d \times d$  matrices with entries  $\pm 1$  and let

$$H = \begin{bmatrix} A & D & B & C \\ -D & A & -C & B \\ -B & C & A & -D \\ -C & -B & D & A \end{bmatrix},$$

i.e.  $H = A \otimes I + B \otimes i + C \otimes j + D \otimes k$ , where the  $4 \times 4$  matrices  $I, i, j, k$  are matrix representations of the unit quaternions. It may be immediately verified that  $H$  is a Hadamard matrix of order  $n = 4d$  if

$$A^t A + B^t B + C^t C + D^t D = 4dI_d$$

and

$$X^t Y = Y^t X$$

for every two distinct matrices  $X, Y$  from the set  $\{A, B, C, D\}$ . The first infinite class of Hadamard matrices of Williamson type was found by Turyn (1972), who showed that they exist for all orders  $n = 2(q + 1)$ , where  $q$  is a prime power and  $q \equiv 1 \pmod{4}$ . Lagrange's theorem that any positive integer is a sum of four squares suggests that Hadamard matrices of Williamson type may exist for all orders  $n \equiv 0 \pmod{4}$ .

The Hadamard matrices constructed by Paley are either symmetric or of the form  $I + S$ , where  $S$  is skew-symmetric. It has been conjectured that in fact Hadamard matrices of both these types exist for all orders  $n \equiv 0 \pmod{4}$ .

### 3 The Art of Weighing

It was observed by Yates (1935) that, if several quantities are to be measured, more accurate results may be obtained by measuring suitable combinations of them than

by measuring each separately. Suppose, for definiteness, that we have  $m$  objects whose weights are to be determined and we perform  $n \geq m$  weighings. The whole experiment may be represented by an  $n \times m$  matrix  $A = (a_{jk})$ . If the  $k$ -th object is not involved in the  $j$ -th weighing, then  $a_{jk} = 0$ ; if it is involved, then  $a_{jk} = +1$  or  $-1$  according as it is placed in the left-hand or right-hand pan of the balance. The individual weights  $\xi_1, \dots, \xi_m$  are connected with the observed results  $\eta_1, \dots, \eta_n$  of the weighings by the system of linear equations

$$y = Ax, \quad (1)$$

where  $x = (\xi_1, \dots, \xi_m)^t \in \mathbb{R}^m$  and  $y = (\eta_1, \dots, \eta_n)^t \in \mathbb{R}^n$ .

We will again denote by  $\|y\|$  the Euclidean norm  $(|\eta_1|^2 + \dots + |\eta_n|^2)^{1/2}$  of the vector  $y$ . Let  $\bar{x} \in \mathbb{R}^m$  have as its coordinates the correct weights and let  $\bar{y} = A\bar{x}$ . If, because of errors of measurement,  $y$  ranges over the ball  $\|y - \bar{y}\| \leq \rho$  in  $\mathbb{R}^n$ , then  $x$  ranges over the ellipsoid  $(x - \bar{x})^t A^t A (x - \bar{x}) \leq \rho^2$  in  $\mathbb{R}^m$ . Since the volume of the ellipsoid is  $[\det(A^t A)]^{-1/2}$  times the volume of the ball, we may regard the best choice of the design matrix  $A$  to be that for which the ellipsoid has minimum volume. Thus we are led to the problem of maximizing  $\det(A^t A)$  among all  $n \times m$  matrices  $A = (a_{jk})$  with  $a_{jk} \in \{0, -1, 1\}$ .

A different approach to the best choice of design matrix leads (by §2) to a similar result. If  $n > m$  the linear system (1) is overdetermined. However, the least squares estimate for the solution of (1) is

$$x = Cy,$$

where  $C = (A^t A)^{-1} A^t$ . Let  $a_k \in \mathbb{R}^n$  be the  $k$ -th column of  $A$  and let  $c_k \in \mathbb{R}^n$  be the  $k$ -th row of  $C$ . Since  $CA = I_m$ , we have  $c_k a_k = 1$ . If  $y$  ranges over the ball  $\|y - \bar{y}\| \leq \rho$  in  $\mathbb{R}^n$ , then  $\xi_k$  ranges over the real interval  $|\xi_k - \bar{\xi}_k| \leq \rho \|c_k\|$ . Thus we may regard the optimal choice of the design matrix  $A$  for measuring  $\xi_k$  to be that for which  $\|c_k\|$  is a minimum.

By Schwarz's inequality (Chapter I, §4),

$$\|c_k\| \|a_k\| \geq 1,$$

with equality only if  $c_k^t$  is a scalar multiple of  $a_k$ . Also  $\|a_k\| \leq n^{1/2}$ , since all elements of  $A$  have absolute value at most 1. Hence  $\|c_k\| \geq n^{-1/2}$ , with equality if and only if all elements of  $a_k$  have absolute value 1 and  $c_k^t = a_k/n$ . It follows that the design matrix  $A$  is optimal for measuring *each* of  $\xi_1, \dots, \xi_m$  if all elements of  $A$  have absolute value 1 and  $A^t A = nI_m$ . Moreover, in this case the least squares estimate for the solution of (1) is simply  $x = A^t y/n$ . Thus the individual weights are easily determined from the observed measurements by additions and subtractions, followed by a division by  $n$ .

Suppose, for example, that  $m = 3$  and  $n = 4$ . If we take

$$A = \begin{bmatrix} + & + & + \\ + & + & - \\ - & + & + \\ + & - & + \end{bmatrix},$$

where  $+$  and  $-$  stand for 1 and  $-1$  respectively, then  $A^t A = 4I_3$ . With this experimental design the individual weights may all be determined with twice the accuracy of the weighing procedure.

The next result shows, in particular, that if we wish to maximize  $\det(A^t A)$  among the  $n \times m$  matrices  $A$  with all entries 0, 1 or  $-1$ , then we may restrict attention to those with all entries 1 or  $-1$ .

**Proposition 8** *Let  $\alpha, \beta$  be real numbers with  $\alpha < \beta$  and let  $\mathcal{S}$  be the set of all  $n \times m$  matrices  $A = (\alpha_{jk})$  such that  $\alpha \leq \alpha_{jk} \leq \beta$  for all  $j, k$ . Then there exists an  $n \times m$  matrix  $M = (\mu_{jk})$  such that  $\mu_{jk} \in \{\alpha, \beta\}$  for all  $j, k$  and*

$$\det(M^t M) = \max_{A \in \mathcal{S}} \det(A^t A).$$

*Proof* For any  $n \times m$  real matrix  $A$ , either the symmetric matrix  $A^t A$  is positive definite and  $\det(A^t A) > 0$ , or  $A^t A$  is positive semidefinite and  $\det(A^t A) = 0$ . Since the result is obvious if  $\det(A^t A) = 0$  for every  $A \in \mathcal{S}$ , we assume that  $\det(A^t A) > 0$  for some  $A \in \mathcal{S}$ . This implies  $m \leq n$ . Partition such an  $A$  in the form

$$A = (vB),$$

where  $v$  is the first column of  $A$  and  $B$  is the remainder. Then

$$A^t A = \begin{bmatrix} v^t v & v^t B \\ B^t v & B^t B \end{bmatrix}$$

and  $B^t B$  is also a positive definite symmetric matrix. By multiplying  $A^t A$  on the left by

$$\begin{bmatrix} I & -v^t B(B^t B)^{-1} \\ 0 & I \end{bmatrix}$$

and taking determinants, we see that

$$\det(A^t A) = f(v) \det(B^t B),$$

where

$$f(v) = v^t v - v^t B(B^t B)^{-1} B^t v.$$

We can write  $f(v) = v^t Q v$ , where

$$Q = I - P, \quad P = B(B^t B)^{-1} B^t.$$

From  $P^t = P = P^2$  we obtain  $Q^t = Q = Q^2$ . Hence  $Q = Q^t Q$  is a positive semidefinite symmetric matrix.

If  $v = \theta v_1 + (1 - \theta)v_2$ , where  $v_1$  and  $v_2$  are fixed vectors and  $\theta \in \mathbb{R}$ , then  $f(v)$  is a quadratic polynomial  $q(\theta)$  in  $\theta$  whose leading coefficient

$$v_1^t Q v_1 - v_2^t Q v_1 - v_1^t Q v_2 + v_2^t Q v_2$$

is nonnegative, since  $Q$  is positive semidefinite. It follows that  $q(\theta)$  attains its maximum value in the interval  $0 \leq \theta \leq 1$  at an endpoint.

Put

$$\mu = \sup_{A \in \mathcal{S}} \det(A^t A).$$

Since  $\det(A^t A)$  is a continuous function of the  $mn$  variables  $\alpha_{jk}$  and  $\mathcal{S}$  may be regarded as a compact set in  $\mathbb{R}^{mn}$ ,  $\mu$  is finite and there exists a matrix  $A \in \mathcal{S}$  for which  $\det(A^t A) = \mu$ . By repeatedly applying the argument of the preceding paragraph to this  $A$  we may replace it by one for which every entry in the first column is either  $\alpha$  or  $\beta$  and for which also  $\det(A^t A) = \mu$ . These operations do not affect the submatrix  $B$  formed by the last  $m - 1$  columns of  $A$ . By interchanging the  $k$ -th column of  $A$  with the first, which does not alter the value of  $\det(A^t A)$ , we may apply the same argument to every other column of  $A$ .  $\square$

The proof of Proposition 8 actually shows that if  $C$  is a compact subset of  $\mathbb{R}^n$  and if  $\mathcal{S}$  is the set of all  $n \times m$  matrices  $A$  whose columns are in  $C$ , then there exists an  $n \times m$  matrix  $M$  whose columns are extreme points of  $C$  such that

$$\det(M^t M) = \sup_{A \in \mathcal{S}} \det(A^t A).$$

Here  $e \in C$  is said to be an *extreme point* of  $C$  if there do not exist distinct  $v_1, v_2 \in C$  and  $\theta \in (0, 1)$  such that  $e = \theta v_1 + (1 - \theta)v_2$ .

The preceding discussion concerns weighings by a chemical balance. If instead we use a spring balance, then we are similarly led to the problem of maximizing  $\det(B^t B)$  among all  $n \times m$  matrices  $B = (\beta_{jk})$  with  $\beta_{jk} = 1$  or  $0$  according as the  $k$ -th object is or is not involved in the  $j$ -th weighing. Moreover other types of measurement lead to the same problem. A spectrometer sorts electromagnetic radiation into bundles of rays, each bundle having a characteristic wavelength. Instead of measuring the intensity of each bundle separately, we can measure the intensity of various combinations of bundles by using masks with open or closed slots.

It will now be shown that in the case  $m = n$  the chemical and spring balance problems are essentially equivalent.

**Lemma 9** *If  $B$  is an  $(n - 1) \times (n - 1)$  matrix of 0's and 1's, and if  $J_n$  is the  $n \times n$  matrix whose entries are all 1, then*

$$A = J_n - \begin{bmatrix} O & O \\ O & 2B \end{bmatrix},$$

*is an  $n \times n$  matrix of 1's and  $-1$ 's, whose first row and column contain only 1's, such that*

$$\det A = (-2)^{n-1} \det B.$$

*Moreover, every  $n \times n$  matrix of 1's and  $-1$ 's, whose first row and column contain only 1's, is obtained in this way.*

*Proof* Since

$$A = \begin{bmatrix} 1 & O \\ e_{n-1}^t & I \end{bmatrix} \begin{bmatrix} 1 & e_{n-1} \\ O & -2B \end{bmatrix},$$

where  $e_m$  denotes a row of  $m$  1's, the matrix  $A$  has determinant  $(-2)^{n-1} \det B$ . The rest of the lemma is obvious.  $\square$

Let  $A$  be an  $n \times n$  matrix with entries  $\pm 1$ . By multiplying rows and columns of  $A$  by  $-1$  we can make all elements in the first row and first column equal to 1 without altering the value of  $\det(A^t A)$ . It follows from Lemma 9 that if  $\alpha_n$  is the maximum of  $\det(A^t A)$  among all  $n \times n$  matrices  $A = (\alpha_{jk})$  with  $\alpha_{jk} \in \{-1, 1\}$ , and if  $\beta_{n-1}$  is the maximum of  $\det(B^t B)$  among all  $(n-1) \times (n-1)$  matrices  $B = (\beta_{jk})$  with  $\beta_{jk} \in \{0, 1\}$ , then

$$\alpha_n = 2^{2n-2} \beta_{n-1}.$$

## 4 Some Matrix Theory

In rectangular coordinates the equation of an ellipse with centre at the origin has the form

$$Q := ax^2 + 2bxy + cy^2 = \text{const.} \quad (*)$$

This is not the form in which the equation of an ellipse is often written, because of the 'cross product' term  $2bxy$ . However, we can bring it to that form by rotating the axes, so that the major axis of the ellipse lies along one coordinate axis and the minor axis along the other. This is possible because the major and minor axes are perpendicular to one another. These assertions will now be verified analytically.

In matrix notation,  $Q = z^t A z$ , where

$$A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}, \quad z = \begin{bmatrix} x \\ y \end{bmatrix}.$$

A rotation of coordinates has the form  $z = T w$ , where

$$T = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}, \quad w = \begin{bmatrix} u \\ v \end{bmatrix}.$$

Then  $Q = w^t B w$ , where  $B = T^t A T$ . Multiplying out, we obtain

$$B = \begin{bmatrix} a' & b' \\ b' & c' \end{bmatrix},$$

where

$$b' = b(\cos^2 \theta - \sin^2 \theta) - (a - c) \sin \theta \cos \theta.$$

To eliminate the cross product term we choose  $\theta$  so that  $b(\cos^2 \theta - \sin^2 \theta) = (a - c) \sin \theta \cos \theta$ ; i.e.,  $2b \cos 2\theta = (a - c) \sin 2\theta$ , or

$$\tan 2\theta = 2b/(a - c).$$

The preceding argument applies equally well to a hyperbola, since it is also described by an equation of the form (\*). We now wish to extend this result to higher dimensions. An  $n$ -dimensional conic with centre at the origin has the form

$$Q := x^t A x = \text{const.},$$

where  $x \in \mathbb{R}^n$  and  $A$  is an  $n \times n$  real symmetric matrix. The analogue of a rotation is a linear transformation  $x = T y$  which preserves Euclidean lengths, i.e.  $x^t x = y^t y$ . This holds for all  $y \in \mathbb{R}^n$  if and only if

$$T^t T = I.$$

A matrix  $T$  which satisfies this condition is said to be *orthogonal*. Then  $T^t = T^{-1}$  and hence also  $T T^t = I$ .

The single most important fact about real symmetric matrices is the *principal axes transformation*:

**Theorem 10** *If  $H$  is an  $n \times n$  real symmetric matrix, then there exists an  $n \times n$  real orthogonal matrix  $U$  such that  $U^t H U$  is a diagonal matrix:*

$$U^t H U = \text{diag}[\lambda_1, \dots, \lambda_n].$$

*Proof* Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be the map defined by

$$f(x) = x^t H x.$$

Since  $f$  is continuous and the unit sphere  $S = \{x \in \mathbb{R}^n : x^t x = 1\}$  is compact,

$$\lambda_1 := \sup_{x \in S} f(x)$$

is finite and there exists an  $x_1 \in S$  such that  $f(x_1) = \lambda_1$ . We are going to show that, if  $x \in S$  and  $x^t x_1 = 0$ , then also  $x^t H x_1 = 0$ .

For any real  $\varepsilon$ , put

$$y = (x_1 + \varepsilon x)/(1 + \varepsilon^2)^{1/2}.$$

Then also  $y \in S$ , since  $x$  and  $x_1$  are orthogonal vectors of unit length. Hence  $f(y) \leq f(x_1)$ , by the definition of  $x_1$ . But  $x_1^t H x = x^t H x_1$ , since  $H$  is symmetric, and hence

$$f(y) = \{f(x_1) + 2\varepsilon x^t H x_1 + \varepsilon^2 f(x)\}/(1 + \varepsilon^2).$$

For small  $|\varepsilon|$  it follows that

$$f(y) = f(x_1) + 2\varepsilon x^t H x_1 + O(\varepsilon^2).$$

If  $x^t H x_1$  were different from zero, we could choose  $\varepsilon$  to have the same sign as it and obtain the contradiction  $f(y) > f(x_1)$ .

On the intersection of the unit sphere  $S$  with the hyperplane  $x^t x_1 = 0$ , the function  $f$  attains its maximum value  $\lambda_2$  at some point  $x_2$ . Similarly, on the intersection of the unit sphere  $S$  with the  $(n-2)$ -dimensional subspace of all  $x$  such that  $x^t x_1 = x^t x_2 = 0$ , the function  $f$  attains its maximum value  $\lambda_3$  at some point  $x_3$ . Proceeding in this way we obtain  $n$  mutually orthogonal unit vectors  $x_1, \dots, x_n$ . Moreover  $x_j^t H x_j = \lambda_j$  and, by the argument of the previous paragraph,  $x_j^t H x_k = 0$  if  $j > k$ . It follows that the matrix  $U$  with columns  $x_1, \dots, x_n$  satisfies all the requirements.  $\square$

It should be noted that, if  $U$  is any orthogonal matrix such that  $U^t H U = \text{diag}[\lambda_1, \dots, \lambda_n]$  then, since  $U U^t = I$ , the columns  $x_1, \dots, x_n$  of  $U$  satisfy

$$H x_j = \lambda_j x_j \quad (1 \leq j \leq n).$$

That is,  $\lambda_j$  is an *eigenvalue* of  $H$  and  $x_j$  a corresponding *eigenvector* ( $1 \leq j \leq n$ ).

A real symmetric matrix  $A$  is *positive definite* if  $x^t A x > 0$  for every real vector  $x \neq 0$  (and *positive semi-definite* if  $x^t A x \geq 0$  for every real vector  $x$  with equality for some  $x \neq 0$ ). It follows from Theorem 10 that two real symmetric matrices can be simultaneously diagonalized, if one of them is positive definite, although the transforming matrix may not be orthogonal:

**Proposition 11** *If  $A$  and  $B$  are  $n \times n$  real symmetric matrices, with  $A$  positive definite, then there exists an  $n \times n$  nonsingular real matrix  $T$  such that  $T^t A T$  and  $T^t B T$  are both diagonal matrices.*

*Proof* By Theorem 10, there exists a real orthogonal matrix  $U$  such that  $U^t A U$  is a diagonal matrix:

$$U^t A U = \text{diag}[\lambda_1, \dots, \lambda_n].$$

Moreover,  $\lambda_j > 0$  ( $1 \leq j \leq n$ ), since  $A$  is positive definite. Hence there exists  $\delta_j > 0$  such that  $\delta_j^2 = 1/\lambda_j$ . If  $D = \text{diag}[\delta_1, \dots, \delta_n]$ , then  $D^t U^t A U D = I$ . By Theorem 10 again, there exists a real orthogonal matrix  $V$  such that

$$V^t (D^t U^t B U D) V = \text{diag}[\mu_1, \dots, \mu_n]$$

is a diagonal matrix. Hence we can take  $T = U D V$ . □

Proposition 11 will now be used to obtain an inequality due to Fischer (1908):

**Proposition 12** *If  $G$  is a positive definite real symmetric matrix, and if*

$$G = \begin{bmatrix} G_1 & G_2 \\ G_2^t & G_3 \end{bmatrix}$$

*is any partition of  $G$ , then*

$$\det G \leq \det G_1 \cdot \det G_3,$$

*with equality if and only if  $G_2 = 0$ .*

*Proof* Since  $G_3$  is also positive definite, we can write  $G = Q^t H Q$ , where

$$Q = \begin{bmatrix} I & 0 \\ G_3^{-1} G_2^t & I \end{bmatrix}, \quad H = \begin{bmatrix} H_1 & 0 \\ 0 & G_3 \end{bmatrix},$$

and  $H_1 = G_1 - G_2 G_3^{-1} G_2^t$ . Since  $\det G = \det H_1 \cdot \det G_3$ , we need only show that  $\det H_1 \leq \det G_1$ , with equality only if  $G_2 = 0$ .

Since  $G_1$  and  $H_1$  are both positive definite, they can be simultaneously diagonalized. Thus, if  $G_1$  and  $H_1$  are  $p \times p$  matrices, there exists a nonsingular real matrix  $T$

such that

$$T^t G_1 T = \text{diag}[\gamma_1, \dots, \gamma_p], \quad T^t H_1 T = \text{diag}[\delta_1, \dots, \delta_p].$$

Since  $G_3^{-1}$  is positive definite,  $u^t (G_1 - H_1) u \geq 0$  for any  $u \in \mathbb{R}^p$ . Hence  $\gamma_i \geq \delta_i > 0$  for  $i = 1, \dots, p$  and  $\det G_1 \geq \det H_1$ . Moreover  $\det G_1 = \det H_1$  only if  $\gamma_i = \delta_i$  for  $i = 1, \dots, p$ .

Hence if  $\det G_1 = \det H_1$ , then  $G_1 = H_1$ , i.e.  $G_2 G_3^{-1} G_2^t = 0$ . Thus  $w^t G_3^{-1} w = 0$  for any vector  $w = G_2^t v$ . Since  $w^t G_3^{-1} w = 0$  implies  $w = 0$ , it follows that  $G_2 = 0$ .  $\square$

From Proposition 12 we obtain by induction

**Proposition 13** *If  $G = (\gamma_{jk})$  is an  $m \times m$  positive definite real symmetric matrix, then*

$$\det G \leq \gamma_{11} \gamma_{22} \cdots \gamma_{mm},$$

*with equality if and only if  $G$  is a diagonal matrix.*

By applying Proposition 13 to the matrix  $G = A^t A$ , we obtain again Proposition 3. Proposition 13 may be sharpened in the following way:

**Proposition 14** *If  $G = (\gamma_{jk})$  is an  $m \times m$  positive definite real symmetric matrix, then*

$$\det G \leq \gamma_{11} \prod_{j=2}^m (\gamma_{jj} - \gamma_{1j}^2 / \gamma_{11}),$$

*with equality if and only if  $\gamma_{jk} = \gamma_{1j} \gamma_{1k} / \gamma_{11}$  for  $2 \leq j < k \leq m$ .*

*Proof* If

$$T = \begin{bmatrix} 1 & g \\ 0 & I_{m-1} \end{bmatrix},$$

where  $g = (-\gamma_{12}/\gamma_{11}, \dots, -\gamma_{1m}/\gamma_{11})$ , then

$$T^t G T = \begin{bmatrix} \gamma_{11} & 0 \\ 0 & H \end{bmatrix},$$

where  $H = (\eta_{jk})$  is an  $(m-1) \times (m-1)$  positive definite real symmetric matrix with entries

$$\eta_{jk} = \gamma_{jk} - \gamma_{1j} \gamma_{1k} / \gamma_{11} \quad (2 \leq j \leq k \leq m).$$

Since  $\det G = \gamma_{11} \det H$ , the result now follows from Proposition 13.  $\square$

Some further inequalities for the determinants of positive definite matrices will now be derived, which will be applied to Hadamard's determinant problem in the next section. We again denote by  $J_m$  the  $m \times m$  matrix whose entries are all 1.

**Lemma 15** If  $C = \alpha I_m + \beta J_m$  for some real  $\alpha, \beta$ , then

$$\det C = \alpha^{m-1}(\alpha + m\beta).$$

Moreover, if  $\det C \neq 0$ , then  $C^{-1} = \gamma I_m + \delta J_m$ , where  $\delta = -\beta\alpha^{-1}(\alpha + m\beta)^{-1}$  and  $\gamma = \alpha^{-1}$ .

*Proof* Subtract the first row of  $C$  from each of the remaining rows, and then add to the first column of the resulting matrix each of the remaining columns. These operations do not alter the determinant and replace  $C$  by an upper triangular matrix with main diagonal entries  $\alpha + m\beta$  (once) and  $\alpha$  ( $m - 1$  times). Hence  $\det C = \alpha^{m-1}(\alpha + m\beta)$ .

If  $\det C \neq 0$  and if  $\gamma, \delta$  are defined as in the statement of the lemma, then from  $J_m^2 = mJ_m$  it follows directly that

$$(\alpha I_m + \beta J_m)(\gamma I_m + \delta J_m) = I_m. \quad \square$$

**Proposition 16** Let  $G = (\gamma_{jk})$  be an  $m \times m$  positive definite real symmetric matrix such that  $|\gamma_{jk}| \geq \beta$  for all  $j, k$  and  $\gamma_{jj} \leq \alpha + \beta$  for all  $j$ , where  $\alpha, \beta > 0$ . Then

$$\det G \leq \alpha^{m-1}(\alpha + m\beta). \quad (2)$$

Moreover, equality holds if and only if there exists a diagonal matrix  $D$ , with main diagonal elements  $\pm 1$ , such that

$$DGD = \alpha I_m + \beta J_m.$$

*Proof* The result is trivial if  $m = 1$  and is easily verified if  $m = 2$ . We assume  $m > 2$  and use induction on  $m$ . By replacing  $G$  by  $DGD$ , where  $D$  is a diagonal matrix whose main diagonal elements have absolute value 1, we may suppose that  $\gamma_{1k} \geq 0$  for  $2 \leq k \leq m$ . Since the determinant is a linear function of its rows, we have

$$\det G = (\gamma_{11} - \beta)\delta + \eta,$$

where  $\delta$  is the determinant of the matrix obtained from  $G$  by omitting the first row and column and  $\eta$  is the determinant of the matrix  $H$  obtained from  $G$  by replacing  $\gamma_{11}$  by  $\beta$ . By the induction hypothesis,

$$\delta \leq \alpha^{m-2}(\alpha + m\beta - \beta).$$

If  $\eta \leq 0$ , it follows that

$$\det G \leq \alpha^{m-1}(\alpha + m\beta - \beta) < \alpha^{m-1}(\alpha + m\beta).$$

Thus we now suppose  $\eta > 0$ . Then  $H$  is positive definite, since the submatrix obtained by omitting the first row and column is positive definite. By Proposition 14,

$$\eta \leq \beta \prod_{j=2}^m (\gamma_{jj} - \gamma_{1j}^2 / \beta),$$

with equality only if  $\gamma_{jk} = \gamma_{1j}\gamma_{1k}/\beta$  for  $2 \leq j < k \leq m$ . Hence  $\eta \leq \alpha^{m-1}\beta$ , with equality only if  $\gamma_{jj} = \alpha + \beta$  for  $2 \leq j \leq m$  and  $\gamma_{jk} = \beta$  for  $1 \leq j < k \leq m$ . Consequently

$$\det G \leq \alpha^{m-1}(\alpha + m\beta - \beta) + \alpha^{m-1}\beta = \alpha^{m-1}(\alpha + m\beta),$$

with equality only if  $G = \alpha I_m + \beta J_m$ .  $\square$

A square matrix will be called a *signed permutation matrix* if each row and column contains only one nonzero entry and this entry is 1 or  $-1$ .

**Proposition 17** *Let  $G = (\gamma_{jk})$  be an  $m \times m$  positive definite real symmetric matrix such that  $\gamma_{jj} \leq \alpha + \beta$  for all  $j$  and either  $\gamma_{jk} = 0$  or  $|\gamma_{jk}| \geq \beta$  for all  $j, k$ , where  $\alpha, \beta > 0$ .*

*Suppose in addition that  $\gamma_{ik} = \gamma_{jk} = 0$  implies  $\gamma_{ij} \neq 0$ . Then*

$$\begin{aligned} \det G &\leq \alpha^{m-2}(\alpha + m\beta/2)^2 && \text{if } m \text{ is even,} \\ \det G &\leq \alpha^{m-2}(\alpha + (m+1)\beta/2)(\alpha + (m-1)\beta/2) && \text{if } m \text{ is odd.} \end{aligned} \quad (3)$$

*Moreover, equality holds if and only if there is a signed permutation matrix  $U$  such that*

$$U^t G U = \begin{bmatrix} L & 0 \\ 0 & M \end{bmatrix},$$

*where*

$$\begin{aligned} L = M &= \alpha I_{m/2} + \beta J_{m/2} && \text{if } m \text{ is even,} \\ L &= \alpha I_{(m+1)/2} + \beta J_{(m+1)/2}, M = \alpha I_{(m-1)/2} + \beta J_{(m-1)/2} && \text{if } m \text{ is odd.} \end{aligned}$$

*Proof* We are going to establish the inequality

$$\det G \leq \alpha^{m-2}(\alpha + s\beta)(\alpha + m\beta - s\beta), \quad (4)$$

where  $s$  is the maximum number of zero elements in any row of  $G$ . Since, as a function of the real variable  $s$ , the quadratic on the right of (4) attains its maximum value for  $s = m/2$ , and has the same value for  $s = (m+1)/2$  as for  $s = (m-1)/2$ , this will imply (3). It will also imply that if equality holds in (3), then  $s = m/2$  if  $m$  is even and  $s = (m+1)/2$  or  $(m-1)/2$  if  $m$  is odd.

For  $m = 2$  it is easily verified that (4) holds. We assume  $m > 2$  and use induction. By performing the same signed permutation on rows and columns, we may suppose that the second row of  $G$  has the maximum number  $s$  of zero elements, and that all nonzero elements of the first row are positive and precede the zero elements. All the hypotheses of the proposition remain satisfied by the matrix  $G$  after this operation.

Let  $s'$  be the number of zero elements in the first row and put  $r' = m - s'$ . As in the proof of Proposition 16, we have

$$\det G = (\gamma_{11} - \beta)\delta + \eta,$$

where  $\delta$  is the determinant of the matrix obtained from  $G$  by omitting the first row and column and  $\eta$  is the determinant of the matrix  $H$  obtained from  $G$  by replacing  $\gamma_{11}$  by  $\beta$ . We partition  $H$  in the form

$$H = \begin{bmatrix} L & N \\ N^t & M \end{bmatrix},$$

where  $L, M$  are square matrices of orders  $r', s'$  respectively. By construction all elements in the first row of  $L$  are positive and all elements in the first row of  $N$  are zero. Furthermore, by the hypotheses of the proposition, all elements of  $M$  have absolute value  $\geq \beta$ .

By the induction hypothesis,

$$\delta \leq \alpha^{m-3}(\alpha + s\beta)(\alpha + m\beta - \beta - s\beta).$$

If  $\eta \leq 0$ , it follows immediately that (4) holds with strict inequality. Thus we now suppose  $\eta > 0$ . Then  $H$  is positive definite and hence, by Fischer's inequality (Proposition 12),  $\eta \leq \det L \cdot \det M$ , with equality only if  $N = 0$ . But, by Proposition 14,

$$\det L \leq \beta \prod_{j=2}^{r'} (\gamma_{jj} - \gamma_{1j}^2 / \beta) \leq \alpha^{r'-1} \beta$$

and, by Proposition 16,

$$\det M \leq \alpha^{s'-1}(\alpha + s'\beta).$$

Hence

$$\det G \leq \alpha^{m-2}(\alpha + s\beta)(\alpha + m\beta - \beta - s\beta) + \alpha^{m-2}\beta(\alpha + s'\beta).$$

Since  $s' \leq s$ , it follows that (4) holds and actually with strict inequality if  $s' \neq s$ .

If equality holds in (4) then, by Proposition 14, we must have  $L = \alpha I_{r'} + \beta J_{r'}$ , and by Proposition 16 after normalization we must also have  $M = \alpha I_{s'} + \beta J_{s'}$ .  $\square$

## 5 Application to Hadamard's Determinant Problem

We have seen that, if  $A$  is an  $n \times m$  real matrix with all entries  $\pm 1$ , then  $\det(A^t A) \leq n^m$ , with strict inequality if  $n > 2$  and  $n$  is not divisible by 4. The question arises, what is the maximum value of  $\det(A^t A)$  in such a case? In the present section we use the results of the previous section to obtain some answers to this question. We consider first the case where  $n$  is odd.

**Proposition 18** *Let  $A = (\alpha_{jk})$  be an  $n \times m$  matrix with  $\alpha_{jk} = \pm 1$  for all  $j, k$ . If  $n$  is odd, then*

$$\det(A^t A) \leq (n-1)^{m-1}(n-1+m).$$

*Moreover, equality holds if and only if  $n \equiv 1 \pmod{4}$  and, after changing the signs of some columns of  $A$ ,*

$$A^t A = (n-1)I_m + J_m.$$

*Proof* We may assume  $\det(A^t A) \neq 0$  and thus  $m \leq n$ . Then  $A^t A = G = (\gamma_{jk})$  is a positive definite real symmetric matrix. For all  $j, k$ ,

$$\gamma_{jk} = \alpha_{1j}\alpha_{1k} + \cdots + \alpha_{nj}\alpha_{nk}$$

is an integer and  $\gamma_{jj} = n$ . Moreover  $\gamma_{jk}$  is odd for all  $j, k$ , being the sum of an odd number of  $\pm 1$ 's. Hence the matrix  $G$  satisfies the hypotheses of Proposition 16 with  $\alpha = n - 1$  and  $\beta = 1$ . Everything now follows from Proposition 16, except for the remark that if equality holds we must have  $n \equiv 1 \pmod{4}$ .

But if  $G = (n - 1)I_m + J_m$ , then  $\gamma_{jk} = 1$  for  $j \neq k$ . It now follows, by the argument used in the proof of Proposition 5, that any two distinct columns of  $A$  have the same entries in exactly  $(n + 1)/2$  rows, and any three distinct columns of  $A$  have the same entries in exactly  $(n + 3)/4$  rows. Thus  $n \equiv 1 \pmod{4}$ .  $\square$

Even if  $n \equiv 1 \pmod{4}$  there is no guarantee that the upper bound in Proposition 18 is attained. However the question may be reduced to the existence of  $H$ -matrices if  $m \neq n$ . For suppose  $m \leq n - 1$  and there exists an  $(n - 1) \times m$   $H$ -matrix  $B$ . If we put

$$A = \begin{bmatrix} B \\ e_m \end{bmatrix},$$

where  $e_m$  again denotes a row of  $m$  1's, then  $A^t A = (n - 1)I_m + J_m$ .

On the other hand if  $m = n$ , then equality in Proposition 18 can hold only under very restrictive conditions. For in this case

$$(\det A)^2 = \det A^t A = (n - 1)^{n-1}(2n - 1)$$

and, since  $n$  is odd, it follows that  $2n - 1$  is the square of an integer. It is an open question whether the upper bound in Proposition 18 is always attained when  $m = n$  and  $2n - 1$  is a square. However the nature of an extremal matrix, if one exists, can be specified rather precisely:

**Proposition 19** *If  $A = (\alpha_{jk})$  is an  $n \times n$  matrix with  $n > 1$  odd and  $\alpha_{jk} = \pm 1$  for all  $j, k$ , then*

$$\det(A^t A) \leq (n - 1)^{n-1}(2n - 1).$$

*Moreover if equality holds, then  $n \equiv 1 \pmod{4}$ ,  $2n - 1 = s^2$  for some integer  $s$  and, after changing the signs of some rows and columns of  $A$ , the matrix  $A$  must satisfy*

$$A^t A = (n - 1)I_n + J_n, \quad A J_n = s J_n.$$

*Proof* By Proposition 18 and the preceding remarks, it only remains to show that if there exists an  $A$  such that  $A^t A = (n - 1)I_n + J_n$  then, by changing the signs of some rows, we can ensure that also  $A J_n = s J_n$ .

Since  $\det(AA^t) = \det(A^t A)$ , it follows from Proposition 18 that there exists a diagonal matrix  $D$  with  $D^2 = I_n$  such that

$$DAA^t D = (n - 1)I_n + J_n = A^t A.$$

Replacing  $A$  by  $DA$ , we obtain  $AA^t = A^tA$ . Then  $A$  commutes with  $A^tA$  and hence also with  $J_n$ . Thus the rows and columns of  $A$  all have the same sum  $s$  and  $AJ_n = sJ_n = A^tJ_n$ . Moreover  $s^2 = 2n - 1$ , since

$$s^2 J_n = s A^t J_n = A^t A J_n = (2n - 1) J_n. \quad \square$$

The maximum value of  $\det(A^t A)$  when  $n \equiv 3 \pmod{4}$  is still a bit of a mystery. We now consider the remaining case when  $n$  is even, but not divisible by 4.

**Proposition 20** *Let  $A = (a_{jk})$  be an  $n \times m$  matrix with  $2 \leq m \leq n$  and  $a_{jk} = \pm 1$  for all  $j, k$ . If  $n \equiv 2 \pmod{4}$  and  $n > 2$ , then*

$$\det(A^t A) \leq (n - 2)^{m-2} (n - 2 + m)^2 \quad \text{if } m \text{ is even,}$$

$$\det(A^t A) \leq (n - 2)^{m-2} (n - 1 + m)(n - 3 + m) \quad \text{if } m \text{ is odd.}$$

Moreover, equality holds if and only if there is a signed permutation matrix  $U$  such that

$$U^t A^t A U = \begin{bmatrix} L & 0 \\ 0 & M \end{bmatrix},$$

where

$$L = M = (n - 2)I_{m/2} + 2J_{m/2} \quad \text{if } m \text{ is even,}$$

$$L = (n - 2)I_{(m+1)/2} + 2J_{(m+1)/2}, \quad M = (n - 2)I_{(m-1)/2} + 2J_{(m-1)/2} \quad \text{if } m \text{ is odd.}$$

*Proof.* We need only show that  $G = A^t A$  satisfies the hypotheses of Proposition 17 with  $\alpha = n - 2$  and  $\beta = 2$ . We certainly have  $\gamma_{jj} = n$ . Moreover all  $\gamma_{jk}$  are even, since  $n$  is even and

$$\gamma_{jk} = \alpha_{1j}\alpha_{1k} + \cdots + \alpha_{nj}\alpha_{nk}.$$

Hence  $|\gamma_{jk}| \geq 2$  if  $\gamma_{jk} \neq 0$ . Finally, if  $j, k, \ell$ , are all different and  $\gamma_{j\ell} = \gamma_{k\ell} = 0$ , then

$$\sum_{i=1}^n (\alpha_{ij} + \alpha_{ik})(\alpha_{ij} + \alpha_{i\ell}) = n + \gamma_{jk}.$$

Since  $n \equiv 2 \pmod{4}$ , it follows that also  $\gamma_{jk} \equiv 2 \pmod{4}$  and thus  $\gamma_{jk} \neq 0$ .  $\square$

Again there is no guarantee that the upper bound in Proposition 20 is attained. However the question may be reduced to the existence of  $H$ -matrices if  $m \neq n, n - 1$ . For suppose  $m \leq n - 2$  and there exists an  $(n - 2) \times m$   $H$ -matrix  $B$ . If we put

$$A = \begin{bmatrix} B \\ C \end{bmatrix},$$

where

$$C = \begin{bmatrix} e_r & e_s \\ e_r & -e_s \end{bmatrix},$$

and  $r + s = m$ , then

$$A^t A = \begin{bmatrix} (n-2)I_r + 2J_r & 0 \\ 0 & (n-2)I_s + 2J_s \end{bmatrix}.$$

Thus the upper bound in Proposition 20 is attained by taking  $r = s = m/2$  when  $m$  is even and  $r = (m+1)/2$ ,  $s = (m-1)/2$  when  $m$  is odd.

Suppose now that  $m = n$  and

$$A^t A = \begin{bmatrix} L & 0 \\ 0 & L \end{bmatrix},$$

where  $L = (n-2)I_{n/2} + 2J_{n/2}$ . If  $B$  is the  $n \times (n-1)$  submatrix of  $A$  obtained by omitting the last column, then

$$B^t B = \begin{bmatrix} L & 0 \\ 0 & M \end{bmatrix},$$

where  $M = (n-2)I_{n/2-1} + 2J_{n/2-1}$ . Thus if the upper bound in Proposition 20 is attained for  $m = n$ , then it is also attained for  $m = n-1$ . Furthermore, since

$$\det(AA^t) = \det(A^t A),$$

it follows from Proposition 20 that there exists a signed permutation matrix  $U$  such that

$$UAA^tU^t = A^tA.$$

Replacing  $A$  by  $UA$ , we obtain  $AA^t = A^tA$ . Then  $A$  commutes with  $A^tA$ . If

$$A = \begin{bmatrix} X & Y \\ Z & W \end{bmatrix},$$

is the partition of  $A$  into square submatrices of order  $n/2$ , it follows that  $X, Y, Z, W$  all commute with  $L$  and hence with  $J_{n/2}$ . This means that the entries in any row or any column of  $X$  have the same sum, which we will denote by  $x$ . Similarly the entries in any row or any column of  $Y, Z, W$  have the same sum, which will be denoted by  $y, z, w$  respectively. We may assume  $x, y, w \geq 0$  by replacing  $A$  by

$$\begin{bmatrix} I_{n/2} & 0 \\ 0 & \pm I_{n/2} \end{bmatrix} A \begin{bmatrix} \pm I_{n/2} & 0 \\ 0 & \pm I_{n/2} \end{bmatrix},$$

We have

$$X^tX + Z^tZ = Y^tY + W^tW = L, \quad X^tY + Z^tW = 0,$$

and

$$XX^t + YY^t = ZZ^t + WW^t = L, \quad XZ^t + YW^t = 0.$$

Postmultiplying by  $J$ , we obtain

$$x^2 + z^2 = y^2 + w^2 = 2n - 2, \quad xy + zw = 0,$$

and

$$x^2 + y^2 = z^2 + w^2 = 2n - 2, \quad xz + yw = 0.$$

Adding, we obtain  $x^2 = w^2$  and hence  $x = w$ . Thus  $z^2 = y^2$  and actually  $z = -y$ , since  $xy + zw = 0$ .

This shows, in particular, that if the upper bound in Proposition 20 is attained for  $m = n \equiv 2 \pmod{4}$ , then  $2n - 2 = x^2 + y^2$ , where  $x$  and  $y$  are integers. By Proposition II.39, such a representation is possible if and only if, for every prime  $p \equiv 3 \pmod{4}$ , the highest power of  $p$  which divides  $n - 1$  is even. Hence the upper bound in Proposition 20 is never attained if  $m = n = 22$ . On the other hand if  $m = n = 6$ , then  $2n - 2 = 10 = 9 + 1$  and an extremal matrix  $A$  is obtained by taking  $W = X = J_3$  and  $Z = -Y = 2I_3 - J_3$ .

It is an open question whether the upper bound in Proposition 20 is always attained when  $m = n$  and  $2n - 2$  is a sum of two squares. It is also unknown if, when an extremal matrix exists, one can always take  $W = X$  and  $Z = -Y$ .

## 6 Designs

A *design* (in the most general sense) is a pair  $(P, \mathcal{B})$ , where  $P$  is a finite set of elements, called *points*, and  $\mathcal{B}$  is a collection of subsets of  $P$ , called *blocks*. If  $p_1, \dots, p_v$  are the points of the design and  $B_1, \dots, B_b$  the blocks, then the *incidence matrix* of the design is the  $v \times b$  matrix  $A = (a_{ij})$  of 0's and 1's defined by

$$a_{ij} = \begin{cases} 1 & \text{if } p_i \in B_j, \\ 0 & \text{if } p_i \notin B_j. \end{cases}$$

Conversely, any  $v \times b$  matrix  $A = (a_{ij})$  of 0's and 1's defines in this way a design. However, two such matrices define the same design if one can be obtained from the other by permutations of the rows and columns.

We will be interested in designs with rather more structure. A *2-design* or, especially in older literature, a 'balanced incomplete block design' (*BIBD*) is a design, with more than one point and more than one block, in which each block contains the same number  $k$  of points, each point belongs to the same number  $r$  of blocks, and every pair of distinct points occurs in the same number  $\lambda$  of blocks.

Thus each column of the incidence matrix contains  $k$  1's and each row contains  $r$  1's. Counting the total number of 1's in two ways, by columns and by rows, we obtain

$$bk = vr.$$

Similarly, by counting in two ways the 1's which lie below the 1's in the first row, we obtain

$$r(k - 1) = \lambda(v - 1).$$

Thus if  $v, k, \lambda$  are given, then  $r$  and  $b$  are determined and we may speak of a  $2-(v, k, \lambda)$  design. Since  $v > 1$  and  $b > 1$ , we have

$$1 < k < v, \quad 1 \leq \lambda < r.$$

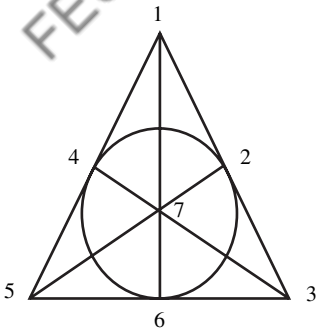


Fig. 1. The Fano plane.

A  $v \times b$  matrix  $A = (\alpha_{ij})$  of 0’s and 1’s is the incidence matrix of a 2-design if and only if, for some positive integers  $k, r, \lambda$ ,

$$\sum_{i=1}^v \alpha_{ij} = k, \quad \sum_{k=1}^b \alpha_{ik}^2 = r, \quad \sum_{k=1}^b \alpha_{ik} \alpha_{jk} = \lambda \quad \text{if } i \neq j \ (1 \leq i, j \leq v),$$

or in other words,

$$e_v A = k e_b, \quad A A^t = (r - \lambda) I_v + \lambda J_v, \tag{5}$$

where  $e_n$  is the  $1 \times n$  matrix with all entries 1,  $I_n$  is the  $n \times n$  unit matrix and  $J_n$  is the  $n \times n$  matrix with all entries 1.

Designs have been used extensively in the design of agricultural and other experiments. To compare the yield of  $v$  varieties of a crop on  $b$  blocks of land, it would be expensive to test each variety separately on each block. Instead we can divide each block into  $k$  plots and use a  $2-(v, k, \lambda)$  design, where  $\lambda = bk(k - 1)/v(v - 1)$ . Then each variety is used exactly  $r = bk/v$  times, no variety is used more than once in any block, and any two varieties are used together in exactly  $\lambda$  blocks. As an example, take  $v = 4, b = 6, k = 2$  and hence  $\lambda = 1, r = 3$ .

Some examples of 2-designs are the finite projective planes. In fact a *projective plane* of order  $n$  may be defined as a  $2-(v, k, \lambda)$  design with

$$v = n^2 + n + 1, \quad k = n + 1, \quad \lambda = 1.$$

It follows that  $b = v$  and  $r = k$ . The blocks in this case are called ‘lines’. The projective plane of order 2, or *Fano plane*, is illustrated in Figure 1. There are seven points and seven blocks, the blocks being the six triples of collinear points and the triple of points on the circle.

Consider now an arbitrary  $2-(v, k, \lambda)$  design. By (5) and Lemma 15,

$$\det(AA^t) = (r - \lambda)^{v-1} (r - \lambda + \lambda v) > 0,$$

since  $r > \lambda$ . This implies the inequality  $b \geq v$ , due to Fisher (1940), since  $AA^t$  would be singular if  $b < v$ .

A 2-design is said to be *square* or (more commonly, but misleadingly) ‘symmetric’ if  $b = v$ , i.e. if the number of blocks is the same as the number of points. Thus any projective plane is a square 2-design.

For a square  $2-(v, k, \lambda)$  design,  $k = r$  and the incidence matrix  $A$  is itself nonsingular. The first relation (5) is now equivalent to  $J_v A = k J_v$ . Since  $k = r$ , the sum of the entries in any row of  $A$  is also  $k$  and thus  $J_v A^t = k J_v$ . By multiplying the second relation (5) on the left by  $A^{-1}$  and on the right by  $A$ , we further obtain

$$A^t A = (r - \lambda)I_v + \lambda J_v.$$

Thus  $A^t$  is also the incidence matrix of a square  $2-(v, k, \lambda)$  design, the *dual* of the given design.

This partly combinatorial argument may be replaced by a more general matrix one:

**Lemma 21** *Let  $a, b, k$  be real numbers and  $n > 1$  an integer. There exists a nonsingular real  $n \times n$  matrix  $A$  such that*

$$AA^t = aI + bJ, \quad JA = kJ, \quad (6)$$

*if and only if  $a > 0$ ,  $a + bn > 0$  and  $k^2 = a + bn$ . Moreover any such matrix  $A$  also satisfies*

$$A^t A = aI + bJ, \quad JA^t = kJ. \quad (7)$$

*Proof* We show first that if  $A$  is any real  $n \times n$  matrix satisfying (6), then  $a + bn = k^2$ . In fact, since  $J^2 = nJ$ , the first relation in (6) implies  $JAA^tJ = (a + bn)nJ$ , whereas the second implies  $JAA^tJ = k^2nJ$ .

We show next that the symmetric matrix  $G := aI + bJ$  is positive definite if and only if  $a > 0$  and  $a + bn > 0$ . By Lemma 15,  $\det G = a^{n-1}(a + bn)$ . If  $G$  is positive definite, its determinant is positive. Since all principal submatrices are also positive definite, we must have  $a^{i-1}(a + bi) > 0$  for  $1 \leq i \leq n$ . In particular,  $a + b > 0$ ,  $a(a + 2b) > 0$ , which is only possible if  $a > 0$ . It now follows that also  $a + bn > 0$ .

Conversely, suppose  $a > 0$  and  $a + bn > 0$ . Then  $\det G > 0$  and there exist nonzero real numbers  $h, k$  such that  $a = h^2$ ,  $a + bn = k^2$ . If we put  $C = hI + (k - h)n^{-1}J$ , then  $JC = kJ$  and

$$C^2 = h^2I + \{2h(k - h) + (k - h)^2\}n^{-1}J = aI + bJ = G.$$

Since  $\det G > 0$ , this shows that  $G = CC^t$  is positive definite and  $C$  is nonsingular.

Finally, let  $A$  be any nonsingular real  $n \times n$  matrix satisfying (6). Since  $A$  is nonsingular,  $AA^t$  is a positive definite symmetric matrix and hence  $a > 0$ ,  $a + bn > 0$ . Since  $AA^t = C^2$  and  $C^t = C$ , we have  $A = CU$ , where  $U$  is orthogonal. Hence  $A^t = U^tC$  and  $C = UA^t$ . From  $JC = kJ$  we obtain  $kJ = JA = JCU = kJU$ . Thus  $J = JU$  and  $JA^t = JUA^t = JC = kJ$ . Moreover  $U^tJU = J$ , since  $J^t = J$ , and hence

$$A^t A = U^t C^2 U = U^t (aI + bJ) U = aI + bJ. \quad \square$$

In Chapter VII we will derive necessary and sufficient conditions for the existence of a nonsingular *rational*  $n \times n$  matrix  $A$  such that  $AA^t = aI + bJ$ , and thus in particular obtain some basic restrictions on the parameters  $v, k, \lambda$  for the existence of a square  $2-(v, k, \lambda)$  design. These were first obtained by Bruck, Ryser and Chowla (1949/50).

We now consider the relationship between designs and Hadamard's determinant problem. By passing from  $A$  to  $B = (J_n - A^t)/2$ , it may be seen immediately that equality holds in Proposition 19 if and only if there exists a  $2$ -( $n, k, \lambda$ ) design, where  $k = (n - s)/2$ ,  $\lambda = (n + 1 - 2s)/4$  and  $s^2 = 2n - 1$ .

We now show that with any Hadamard matrix  $A = (\alpha_{jk})$  of order  $n = 4d$  there is associated a  $2$ -( $4d - 1, 2d - 1, d - 1$ ) design. Assume without loss of generality that all elements in the first row and column of  $A$  are 1. We take  $P = \{2, \dots, n\}$  as the set of points and  $\mathcal{B} = \{B_2, \dots, B_n\}$  as the set of blocks, where  $B_k = \{j \in P : \alpha_{jk} = 1\}$ . Then  $B_k$  has cardinality  $|B_k| = n/2 - 1$  for  $k = 2, \dots, n$ . Moreover, if  $T$  is any subset of  $P$  with  $|T| = 2$ , then the number of blocks containing  $T$  is  $n/4 - 1$ . The argument may also be reversed to show that any  $2$ -( $4d - 1, 2d - 1, d - 1$ ) design is associated in this way with a Hadamard matrix of order  $4d$ .

In particular, for  $d = 2$ , the  $2$ -( $7, 3, 1$ ) design associated with the Hadamard matrix  $H_2 \otimes H_2 \otimes H_2$ , where

$$H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix},$$

is the projective plane of order 2 (Fano plane) illustrated in Figure 1.

The connection between Hadamard matrices and designs may also be derived by a matrix argument. If

$$A = \begin{bmatrix} 1 & e_{n-1} \\ e_{n-1}^t & A \end{bmatrix},$$

is a Hadamard matrix of order  $n = 4d$ , normalized so that its first row and column contain only 1's, then  $B = (J_{n-1} + A)/2$  is a matrix of 0's and 1's such that

$$J_{4d-1}B = (2d - 1)J_{4d-1}, \quad BB^t = dI_{4d-1} + (d - 1)J_{4d-1}.$$

The optimal spring balance design of order  $4d - 1$ , which is obtained by taking  $C = (J_{n-1} - A)/2$ , is a  $2$ -( $4d - 1, 2d, d$ ) design, since

$$J_{4d-1}C = 2dJ_{4d-1}, \quad CC^t = dI_{4d-1} + dJ_{4d-1}.$$

The notion of 2-design will now be generalized. Let  $t, v, k, \lambda$  be positive integers with  $v \geq k \geq t$ . A  $t$ -( $v, k, \lambda$ ) design, or simply a  $t$ -design, is a pair  $(P, \mathcal{B})$ , where  $P$  is a set of cardinality  $v$  and  $\mathcal{B}$  is a collection of subsets of  $P$ , each of cardinality  $k$ , such that any subset of  $P$  of cardinality  $t$  is contained in exactly  $\lambda$  elements of  $\mathcal{B}$ . The elements of  $P$  will be called *points* and the elements of  $\mathcal{B}$  will be called *blocks*. A  $t$ -( $v, k, \lambda$ ) design with  $\lambda = 1$  is known as a *Steiner system*. The *automorphism group* of a  $t$ -design is the group of all permutations of the points which map blocks onto blocks.

If  $t = 1$ , then each point is contained in exactly  $\lambda$  blocks and so the number of blocks is  $\lambda v/k$ . Suppose now that  $t > 1$ . Let  $S$  be a fixed subset of  $P$  of cardinality  $t - 1$  and let  $\lambda'$  be the number of blocks which contain  $S$ . Consider the number of pairs  $(T, B)$ , where  $B \in \mathcal{B}$ ,  $S \subseteq T \subseteq B$  and  $|T| = t$ . By first fixing  $B$  and varying  $T$  we see that this number is  $\lambda'(k - t + 1)$ . On the other hand, by first fixing  $T$  and varying  $B$  we see that this number is  $\lambda(v - t + 1)$ . Hence

$$\lambda' = \lambda(v - t + 1)/(k - t + 1)$$

does not depend on the choice of  $S$  and a  $t$ -( $v, k, \lambda$ ) design  $(P, \mathcal{B})$  is also a  $(t - 1)$ -( $v, k, \lambda'$ ) design. By repeating this argument, we see that each point is contained in exactly  $r$  blocks, where

$$r = \lambda(v - t + 1) \cdots (v - 1)/(k - t + 1) \cdots (k - 1),$$

and the total number of blocks is  $b = rv/k$ . In particular, any  $t$ -design with  $t > 2$  is also a 2-design.

With any Hadamard matrix  $A = (\alpha_{jk})$  of order  $n = 4d$  there is, in addition, associated a  $3$ -( $4d, 2d, d - 1$ ) design. For assume without loss of generality that all elements in the first column of  $A$  are 1. We take  $P = \{1, 2, \dots, n\}$  as the set of points and  $\{B_2, \dots, B_n, B'_2, \dots, B'_n\}$  as the set of blocks, where  $B_k = \{j \in P : \alpha_{jk} = 1\}$  and  $B'_k = \{j \in P : \alpha_{jk} = -1\}$ . Then, by Proposition 5,  $|B_k| = |B'_k| = n/2$  for  $k = 2, \dots, n$ . If  $T$  is any subset of  $P$  with  $|T| = 3$ , say  $T = \{i, j, \ell\}$ , then the number of blocks containing  $T$  is the number of  $k > 1$  such that  $\alpha_{ik} = \alpha_{jk} = \alpha_{\ell k}$ . But, by Proposition 5 again, the number of columns of  $A$  which have the same entries in rows  $i, j, \ell$  is  $n/4$  and this includes the first column. Hence  $T$  is contained in exactly  $n/4 - 1$  blocks. Again the argument may be reversed to show that any  $3$ -( $4d, 2d, d - 1$ ) design is associated in this way with a Hadamard matrix of order  $4d$ .

## 7 Groups and Codes

A group is said to be *simple* if it contains more than one element and has no *normal* subgroups besides itself and the subgroup containing only the identity element. The finite simple groups are in some sense the building blocks from which all finite groups are constructed. There are several infinite families of them: the cyclic groups  $C_p$  of prime order  $p$ , the alternating groups  $\mathcal{A}_n$  of all even permutations of  $n$  objects ( $n \geq 5$ ), the groups  $PSL_n(q)$  derived from the general linear groups of all invertible linear transformations of an  $n$ -dimensional vector space over a finite field of  $q = p^m$  elements ( $n \geq 2$  and  $q > 3$  if  $n = 2$ ), and some other families similar to the last which are analogues for a finite field of the simple Lie groups.

In addition to these infinite families there are 26 *sporadic* finite simple groups. (The *classification theorem* states that there are no other finite simple groups besides those already mentioned. The proof of the classification theorem at present occupies thousands of pages, scattered over a variety of journals, and some parts are actually still unpublished.) All except five of the sporadic groups were found in the years 1965–1981. However, the first five were found by Mathieu (1861, 1873):  $M_{12}$  is a 5-fold transitive group of permutations of 12 objects of order  $12 \cdot 11 \cdot 10 \cdot 9 \cdot 8$  and  $M_{11}$  the subgroup of all permutations in  $M_{12}$  which fix one of the objects;  $M_{24}$  is a 5-fold transitive group of permutations of 24 objects of order  $24 \cdot 23 \cdot 22 \cdot 21 \cdot 20 \cdot 48$ ,  $M_{23}$  the subgroup of all permutations in  $M_{24}$  which fix one of the objects and  $M_{22}$  the subgroup of all permutations which fix two of the objects. The Mathieu groups may be defined in several ways, but the definitions by means of Hadamard matrices that we are going to give are certainly competitive with the others.

Two  $n \times n$  Hadamard matrices  $H_1, H_2$  are said to be *equivalent* if one may be obtained from the other by interchanging two rows or two columns, or by changing the sign of a row or a column, or by any finite number of such operations. Otherwise expressed,  $H_2 = PH_1Q$ , where  $P$  and  $Q$  are signed permutation matrices. An *automorphism* of a Hadamard matrix  $H$  is an equivalence of  $H$  with itself:  $H = PHQ$ . Since  $P = HQ^{-1}H^{-1}$ , the automorphism is uniquely determined by  $Q$ . Under matrix multiplication all admissible  $Q$  form a group  $\mathcal{G}$ , the *automorphism group* of the Hadamard matrix  $H$ . Evidently  $-I \in \mathcal{G}$  and  $-I$  commutes with all elements of  $\mathcal{G}$ . The factor group  $\mathcal{G}/\{\pm I\}$ , obtained by identifying  $Q$  and  $-Q$ , may be called the *reduced automorphism group* of  $H$ .

To illustrate these concepts we will show that all Hadamard matrices of order 12 are equivalent. In fact rather more is true:

**Proposition 22** *Any Hadamard matrix of order 12 may be brought to the form*

$$\begin{array}{cccc}
 + & + & + & + \\
 + & + & + & - \\
 + & - & - & - \\
 + & - & + & - \\
 + & + & - & - \\
 - & + & + & + \\
 + & - & + & - \\
 + & - & + & - \\
 + & + & - & - \\
 - & + & + & - \\
 + & + & - & - \\
 - & + & + & -
 \end{array}
 \begin{array}{cccc}
 + & + & + & + \\
 + & + & + & - \\
 - & - & - & - \\
 + & + & + & - \\
 - & + & - & - \\
 - & - & + & - \\
 - & - & + & - \\
 - & - & + & - \\
 - & - & + & - \\
 - & - & + & - \\
 - & - & + & - \\
 - & - & + & -
 \end{array}
 \begin{array}{cccc}
 + & + & + & + \\
 - & - & - & - \\
 + & + & + & - \\
 - & + & - & - \\
 - & - & + & - \\
 + & + & - & - \\
 + & - & - & + \\
 - & - & + & - \\
 - & - & + & - \\
 - & - & + & - \\
 - & - & + & - \\
 - & - & + & -
 \end{array}
 \begin{array}{cccc}
 + & + & + & + \\
 - & - & - & - \\
 - & - & - & - \\
 + & - & - & + \\
 + & - & - & + \\
 + & - & - & + \\
 + & - & - & + \\
 + & - & - & + \\
 + & - & - & + \\
 + & - & - & + \\
 + & - & - & + \\
 + & - & - & +
 \end{array}
 \begin{array}{c}
 (*)
 \end{array}$$

(where  $+$  stands for 1 and  $-$  for  $-1$ ) by changing the signs of some rows and columns, by permuting the columns, and by permuting the first three rows and the last seven rows.

*Proof* Let  $A = (a_{jk})$  be a Hadamard matrix of order 12. By changing the signs of some columns we may assume that all elements of the first row are  $+1$ . Then, by the orthogonality relations, half the elements of any other row are  $+1$ . By permuting the columns we may assume that all elements in the first half of the second row are  $+1$ . It now follows from the orthogonality relations that in any row after the second the sum of all elements in each half is zero. Hence, by permuting the columns within each half we may assume that the third row is the same as the third row of the array  $(*)$  displayed above. In the  $r$ -th row, where  $r > 3$ , let  $\rho_k$  be the sum of the entries in the  $k$ -th block of three columns ( $k = 1, 2, 3, 4$ ). The orthogonality relations now imply that

$$\rho_1 = \rho_4 = -\rho_2 = -\rho_3.$$

In the  $s$ -th row, where  $s > 3$  and  $s \neq r$ , let  $\sigma_k$  be the sum of the entries in the  $k$ -th block of three columns. Then also

$$\sigma_1 = \sigma_4 = -\sigma_2 = -\sigma_3.$$

If  $\rho_1 = \pm 3$ , then all elements of the same triple of columns in the  $r$ -th row have the same sign and orthogonality to the  $s$ -th row implies  $\sigma_1 = 0$ , which is impossible because  $\sigma_1$  is odd. Hence  $\rho_1 = \pm 1$ . By changing the signs of some rows we may assume that  $\rho_1 = 1$  for every  $r > 3$ . By permuting columns within each block of three we may also normalize the 4-th row, so that the first four rows are now the same as the first four rows of the array (\*).

In any row after the third, within a given block of three columns two elements have the same sign and the third element the opposite sign. Moreover, these signs depend only on the block and not on the row, since  $\rho_1 = 1$ . The scalar product of the triples from two different rows belonging to the same block of columns is 3 if the exceptional elements have the same position in the triple and is  $-1$  otherwise. Since the two rows are orthogonal, the exceptional elements must have the same position in exactly one of the four blocks of columns. Thus if two rows after the 4-th have the same triple of elements in the  $k$ -th block as the 4-th row, then they have no other triple in common with the 4-th row or with one another. But this implies that if one of the two rows is given, then the other is uniquely determined. Hence no other row besides these two has the same triple of elements in the  $k$ -th block as the 4-th row. Since there are eight rows after the 4-th, and since each has exactly one triple in common with the 4-th row, it follows that, for each  $k \in \{1, 2, 3, 4\}$ , exactly two of them have the same triple in the  $k$ -th block as the 4-th row.

The first four rows are unaltered by the following operations:

- (i) interchange of the first and last columns of any triple of columns,
- (ii) interchange of the second and third triple of columns, and then interchange of the second and third rows,
- (iii) interchange of the first and fourth triple of columns, then interchange of the second and third rows and change of sign of these two rows,
- (iv) interchange of the second and fourth triple of columns and change of their signs, then interchange of the first and third rows.

If we denote the elements of the  $r$ -th row ( $r > 4$ ) by  $\zeta_1, \dots, \zeta_{12}$ , then we have

$$\zeta_1 + \zeta_2 + \zeta_3 = 1 = \zeta_{10} + \zeta_{11} + \zeta_{12},$$

$$\zeta_4 + \zeta_5 + \zeta_6 = -1 = \zeta_7 + \zeta_8 + \zeta_9,$$

$$\zeta_2 - \zeta_5 - \zeta_8 + \zeta_{11} = 2.$$

In particular in the 5-th row we have  $\alpha_{52} - \alpha_{55} - \alpha_{58} + \alpha_{5,11} = 2$ . Thus  $\alpha_{52}$  and  $\alpha_{5,11}$  cannot both be  $-1$  and by an operation (iii) we may assume that  $\alpha_{52} = 1$ . Similarly  $\alpha_{55}$  and  $\alpha_{58}$  cannot both be 1 and by an operation (ii) we may assume that  $\alpha_{58} = -1$ . Then  $\alpha_{55} = \alpha_{5,11}$  and by an operation (iv) we may assume that  $\alpha_{55} = \alpha_{5,11} = -1$ . By operations (i) we may finally assume that the 5-th row is the same as the 5-th row of the array (\*).

As we have already shown, exactly one row after the 5-th row has the same triple  $+ - +$  in the last block of columns as the 4-th and 5-th rows and this row must be the same as the 6-th row of the array (\*). By permuting the last seven rows we may assume that this row is also the 6-th row of the given matrix, that the 7-th and 8-th rows have the same first triple of elements as the 4-th row, that the 9-th and 10-th rows have

the same second triple of elements as the 4-th row, and that the 11-th and 12-th rows have the same third triple of elements as the 4-th row.

In any row after the 6-th we have, in addition to the relations displayed above,  $\xi_{11} = 1$ ,  $\xi_{10} + \xi_{12} = 0$  and

$$\xi_1 - \xi_4 - \xi_7 = \xi_2 - \xi_5 - \xi_8 = \xi_3 - \xi_6 - \xi_9 = 1.$$

In the 7-th and 8-th rows we have  $\xi_1 = \xi_3 = 1$ ,  $\xi_2 = -1$ , and hence  $\xi_5 = \xi_8 = -1$ ,  $\xi_4 = -\xi_6 = -\xi_7 = \xi_9$ . Since the first six rows are still unaltered by an operation (ii), and also by interchanging the first and third columns of the last block, we may assume that  $\alpha_{74} = -1$ ,  $\alpha_{7,10} = 1$ . The 7-th and 8-th rows are now uniquely determined and are the same as the 7-th and 8-th rows of the array (\*).

In any row after the 8-th we have

$$\xi_2 - \xi_6 - \xi_7 + \xi_{12} = 2 = \xi_2 - \xi_4 - \xi_9 + \xi_{10}.$$

In the 9-th and 10-th rows we have  $\xi_5 = \xi_{11} = 1$  and  $\xi_4 = \xi_6 = -1$ . Hence  $\xi_2 = -\xi_8 = 1$ ,  $\xi_1 = \xi_7 = -\xi_3 = -\xi_9$ , and finally  $\xi_9 = \xi_{10} = -\xi_{12}$ . Thus the 9-th and 10-th rows are together uniquely determined and may be ordered so as to coincide with the corresponding rows of the array (\*). Similarly the 11-th and 12-th rows are together uniquely determined and may be ordered so as to coincide with the corresponding rows of the displayed array.  $\square$

It follows from Proposition 22 that, for any five distinct rows of a Hadamard matrix of order 12, there exists exactly one pair of columns which either agree in all these rows or disagree in all these rows. Indeed, by permuting the rows we may arrange that the five given rows are the first five rows. Now, by Proposition 22, we may assume that the matrix has the form (\*). But it is evident that in this case there is exactly one pair of columns which either agree or disagree in all the first five rows, namely the 10-th and 12-th columns.

Hence a 5-(12, 6, 1) design is obtained by taking the points to be elements of the set  $P = \{1, \dots, 12\}$  and the blocks to be the  $12 \cdot 11$  subsets  $B_{jk}, B'_{jk}$  with  $j, k \in P$  and  $j \neq k$ , where

$$B_{jk} = \{i \in P : \alpha_{ij} = \alpha_{ik}\}, \quad B'_{jk} = \{i \in P : \alpha_{ij} \neq \alpha_{ik}\}.$$

The Mathieu group  $M_{12}$  may be defined as the automorphism group of this design or as the reduced automorphism group of any Hadamard matrix of order 12.

It is certainly not true in general that all Hadamard matrices of the same order  $n$  are equivalent. For example, there are 60 equivalence classes of Hadamard matrices of order 24. The Mathieu group  $M_{24}$  is connected with the Hadamard matrix of order 24 which is constructed by Paley's method, described in §2. The connection is not as immediate as for  $M_{12}$ , but the ideas involved are of general significance, as we now explain.

A sequence  $x = (\xi_1, \dots, \xi_n)$  of  $n$  0's and 1's may be regarded as a vector in the  $n$ -dimensional vector space  $V = \mathbb{F}_2^n$  over the field of two elements. If we define the weight  $|x|$  of the vector  $x$  to be the number of nonzero coordinates  $\xi_k$ , then

- (i)  $|x| \geq 0$  with equality if and only if  $x = 0$ ,

(ii)  $|x + y| \leq |x| + |y|$ .

The vector space  $V$  acquires the structure of a metric space if we define the (*Hamming distance*) between the vectors  $x$  and  $y$  to be  $d(x, y) = |x - y|$ .

A *binary linear code* is a subspace  $U$  of the vector space  $V$ . If  $U$  has dimension  $k$ , then a *generator matrix* for the code is a  $k \times n$  matrix  $G$  whose rows form a basis for  $U$ . The *automorphism group* of the code is the group of all permutations of the  $n$  coordinates which map  $U$  onto itself. An  $[n, k, d]$ -*binary code* is one for which  $V$  has dimension  $n$ ,  $U$  has dimension  $k$  and  $d$  is the least weight of any nonzero vector in  $U$ .

There are useful connections between codes and designs. Corresponding to any design with incidence matrix  $A$  there is the binary linear code generated over  $\mathbb{F}_2$  by the rows of  $A$ . Given a binary linear code  $U$ , on the other hand, a theorem of Assmus and Mattson (1969) provides conditions under which the nonzero vectors in  $U$  with minimum weight form the rows of the incidence matrix of a  $t$ -design.

Suppose now that  $H$  is a Hadamard matrix of order  $n$ , normalized so that all elements in the first row are 1. Then  $A = (H + J_n)/2$  is a matrix of 0's and 1's with all elements in the first row 1. The code  $C(H)$  defined by the Hadamard matrix  $H$  is the subspace generated by the rows of  $A$ , considered as vectors in the  $n$ -dimensional vector space  $V = \mathbb{F}_2^n$ .

In particular, take  $H = H_{24}$  to be the Hadamard matrix of order 24 formed by Paley's construction:

$$H_{24} = I_{24} + \begin{bmatrix} 0 & e_{23} \\ -e_{23}^t & Q \end{bmatrix},$$

where  $Q = (q_{jk})$  with  $q_{jk} = 0$  if  $j = k$  and otherwise  $= 1$  or  $-1$  according as  $j - k$  is or is not a square mod 23 ( $0 \leq j, k \leq 22$ ). It may be shown that the *extended binary Golay code*  $G_{24} = C(H_{24})$  is a 12-dimensional subspace of  $\mathbb{F}_2^{24}$ , that the minimum weight of any nonzero vector in  $G_{24}$  is 8, and that the sets of nonzero coordinates of the vectors  $x \in G_{24}$  with  $|x| = 8$  form the blocks of a 5-(24, 8, 1) design. The Mathieu group  $M_{24}$  may be defined as the automorphism group of this design or as the automorphism group of the code  $G_{24}$ .

Again, suppose that  $H^{(m)}$  is the Hadamard matrix of order  $n = 2^m$  defined by

$$H^{(m)} = H_2 \otimes \cdots \otimes H_2 \quad (m \text{ factors}),$$

where

$$H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

The *first-order Reed-Muller code*  $R(1, m) = C(H^{(m)})$  is an  $(m+1)$ -dimensional subspace of  $\mathbb{F}_2^n$  and the minimum weight of any nonzero vector in  $R(1, m)$  is  $2^{m-1}$ . It may be mentioned that the 3-( $2^m, 2^{m-1}, 2^{m-2} - 1$ ) design associated with the Hadamard matrix  $H^{(m)}$  has a simple geometrical interpretation. Its points are the points of  $m$ -dimensional affine space over the field of two elements, and its blocks are the hyperplanes of this space (not necessarily containing the origin).

In electronic communication a message is sent as a sequence of 'bits' (an abbreviation for *binary digits*), which may be realised physically by *off* or *on* and which may

be denoted mathematically by 0 or 1. On account of noise the message received may differ slightly from that transmitted, and in some situations it is extremely important to detect and correct the errors. One way of doing so would be to send the same message many times, but it is an inefficient way. Instead suppose the message is composed of *codewords* of length  $n$ , taken from a subspace  $U$  of the vector space  $V = \mathbb{F}_2^n$ . There are  $2^k$  different codewords, where  $k$  is the dimension of  $U$ . If the minimum weight of any nonzero vector in  $U$  is  $d$ , then any two distinct codewords differ in at least  $d$  places. Hence if a codeword  $u \in U$  is transmitted and the received vector  $v \in V$  contains less than  $d/2$  errors, then  $v$  will be closer to  $u$  than to any other codeword. Thus if we are confident that any transmitted codeword will contain less than  $d/2$  errors, we can correct them all by replacing each received vector by the codeword nearest to it.

The Golay code and the first-order Reed–Muller codes are of considerable practical importance in this connection. For the first-order Reed–Muller codes there is a fast algorithm for finding the nearest codeword to any received vector. Photographs of Mars taken by the Mariner 9 spacecraft were transmitted to Earth, using the code  $R(1, 5)$ .

Other *error-correcting codes* are used with compact discs to ensure high quality sound reproduction by eliminating imperfections due, for example, to dust particles.

## 8 Further Remarks

Kowalewski [22] gives a useful traditional account of determinants. Muir [28] is a storehouse of information on special types of determinants; the early Japanese work is described in Mikami [27].

Another approach to determinants, based on the work of Grassmann (1844), should be mentioned here, as it provides easy access to their formal properties and is used in the theory of differential forms. If  $V$  is an  $n$ -dimensional vector space over a field  $F$ , then there exists an associative algebra  $E$ , of dimension  $2^n$  as a vector space over  $F$ , such that

- (a)  $V \subseteq E$ ,
- (b)  $v^2 = 0$  for every  $v \in V$ ,
- (c)  $V$  generates  $E$ , i.e. each element of  $E$  can be expressed as a sum of a scalar multiple of the unit element 1 and of a finite number of products of elements of  $V$ .

The associative algebra  $E$ , which is uniquely determined by these properties, is called the *Grassmann algebra* or *exterior algebra* of the vector space  $V$ . It is easily seen that any two products of  $n$  elements of  $V$  differ only by a scalar factor. Hence, for any linear transformation  $A: V \rightarrow V$ , there exists  $d(A) \in F$  such that

$$(Av_1) \cdots (Av_n) = d(A)v_1 \cdots v_n \quad \text{for all } v_1, \dots, v_n \in V.$$

Evidently  $d(AB) = d(A)d(B)$  and in fact  $d(A) = \det A$ , if we identify  $A$  with its matrix with respect to some fixed basis of  $V$ . This approach to determinants is developed in Bourbaki [6]; see also Barnabei *et al.* [4].

Dieudonné (1943) has extended the notion of determinant to matrices with entries from a division ring; see Artin [1] and Cohn [9]. For a very different method, see Gelfand and Retakh [13].

Hadamard's original paper of 1893 is reproduced in [16]. Surveys on Hadamard matrices have been given by Hedayat and Wallis [19], Seberry and Yamada [34], and Craigen and Wallis [11]. Weighing designs are treated in Raghavarao [31]. For applications of Hadamard matrices to spectrometry, see Harwit and Sloane [18]. The proof of Proposition 8 is due to Shahriari [35].

Our proof of Theorem 10 is a pure existence proof. A more constructive approach was proposed by Jacobi (1846). If one applies to  $n \times n$  matrices the method which we used for  $2 \times 2$  matrices, one can annihilate a symmetric pair of off-diagonal entries. By choosing at each step an off-diagonal pair with maximum absolute value, one obtains a sequence of orthogonal transforms of the given symmetric matrix which converges to a diagonal matrix.

Calculating the eigenvalues of a real symmetric matrix has important practical applications, e.g. to problems of small oscillations in dynamical systems. Householder [21] and Golub and van Loan [14] give accounts of the various computational methods available.

Gantmacher [12] and Horn and Johnson [20] give general treatments of matrix theory, including the inequalities of Hadamard and Fischer. Our discussion of the Hadamard determinant problem for matrices of order not divisible by 4 is mainly based on Wojtas [37]. Further references are given in Neubauer and Ratcliffe [29].

Results of Brouwer (1983) are used in [29] to show that the upper bound in Proposition 19 is attained for infinitely many values of  $n$ . It follows that the upper bound in Proposition 20, with  $m = n$ , is also attained for infinitely many values of  $n$ . For if the  $n \times n$  matrix  $A$  satisfies

$$A^t A = (n-1)I_n + J_n,$$

then the  $2n \times 2n$  matrix

$$\bar{A} = \begin{bmatrix} A & A \\ A & -A \end{bmatrix}$$

satisfies

$$\tilde{A}^t \tilde{A} = \begin{bmatrix} L & O \\ O & L \end{bmatrix},$$

where  $L = 2A^t A = (2n-2)I_n + 2J_n$ .

There are introductions to design theory in Ryser [33], Hall [17], and van Lint and Wilson [25]. For more detailed information, see Brouwer [7], Lander [23] and Beth *et al.* [5]. Applications of design theory are treated in Chapter XIII of [5].

We mention two interesting results which are proved in Chapter 16 of Hall [17]. Given positive integers  $v, k, \lambda$  with  $\lambda < k < v$ :

- (i) If  $k(k-1) = \lambda(v-1)$  and if there exists a  $v \times v$  matrix  $A$  of rational numbers such that

$$AA^t = (k-\lambda)I + \lambda J,$$

then  $A$  may be chosen so that in addition  $JA = kJ$ .

(ii) If there exists a  $v \times v$  matrix  $A$  of integers such that

$$AA^t = (k - \lambda)I + \lambda J, JA = kJ,$$

then every entry of  $A$  is either 0 or 1, and thus  $A$  is the incidence matrix of a square 2-design.

For introductions to the classification theorem for finite simple groups, see Aschbacher [2] and Gorenstein [15]. Detailed information about the finite simple groups is given in Conway *et al.* [10]. There is a remarkable connection between the largest sporadic simple group, nicknamed the 'Monster', and modular forms; see Ray [32].

Good introductions to coding theory are given by van Lint [24] and Pless [30]. MacWilliams and Sloane [26] is more comprehensive, but less up-to-date. Assmus and Mattson [3] is a useful survey article. Connections between codes, designs and graphs are treated in Cameron and van Lint [8]. The historical account in Thompson [36] recaptures the excitement of scientific discovery.

## 9 Selected References

- [1] E. Artin, *Geometric algebra*, reprinted, Wiley, New York, 1988. [Original edition, 1957]
- [2] M. Aschbacher, The classification of the finite simple groups, *Math. Intelligencer* **3** (1980/81), 59–65.
- [3] E.F. Assmus Jr. and H.F. Mattson Jr., Coding and combinatorics, *SIAM Rev.* **16** (1974), 349–388.
- [4] M. Barnabei, A. Brini and G.-C. Rota, On the exterior calculus of invariant theory, *J. Algebra* **96** (1985), 120–160.
- [5] T. Beth, D. Jungnickel and H. Lenz, *Design theory*, 2nd ed., 2 vols., Cambridge University Press, 1999.
- [6] N. Bourbaki, *Algebra I, Chapters 1–3*, Hermann, Paris, 1974. [French original, 1948]
- [7] A.E. Brouwer, Block designs, *Handbook of combinatorics* (ed. R.L. Graham, M. Grötschel and L. Lovász), Vol. I, pp. 693–745, Elsevier, Amsterdam, 1995.
- [8] P.J. Cameron and J.H. van Lint, *Designs, graphs, codes and their links*, Cambridge University Press, 1991.
- [9] P.M. Cohn, *Algebra*, 2nd ed., Vol. 3, Wiley, Chichester, 1991.
- [10] J.H. Conway *et al.*, *Atlas of finite groups*, Clarendon Press, Oxford, 1985.
- [11] R. Craigen and W.D. Wallis, Hadamard matrices: 1893–1993, *Congr. Numer.* **97** (1993), 99–129.
- [12] F.R. Gantmacher, *The theory of matrices*, English transl. by K. Hirsch, 2 vols., Chelsea, New York, 1960.
- [13] I.M. Gelfand and V.S. Retakh, A theory of noncommutative determinants and characteristic functions of graphs, *Functional Anal. Appl.* **26** (1992), 231–246.
- [14] G.H. Golub and C.F. van Loan, *Matrix computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
- [15] D. Gorenstein, Classifying the finite simple groups, *Bull. Amer. Math. Soc. (N.S.)* **14** (1986), 1–98.
- [16] J. Hadamard, Résolution d'une question relative aux déterminants, *Selecta*, pp. 136–142, Gauthier-Villars, Paris, 1935 and *Oeuvres*, Tome I, pp. 239–245, CNRS, Paris, 1968.
- [17] M. Hall, *Combinatorial theory*, 2nd ed., Wiley, New York, 1986.

- [18] M. Harwit and N.J.A. Sloane, *Hadamard transform optics*, Academic Press, New York, 1979.
- [19] A. Hedayat and W.D. Wallis, Hadamard matrices and their applications, *Ann. Statist.* **6** (1978), 1184–1238.
- [20] R.A. Horn and C.A. Johnson, *Matrix analysis*, Cambridge University Press, 1985.
- [21] A.S. Householder, *The theory of matrices in numerical analysis*, Blaisdell, New York, 1964.
- [22] G. Kowalewski, *Einführung in die Determinantentheorie*, 4th ed., de Gruyter, Berlin, 1954.
- [23] E.S. Lander, *Symmetric designs: an algebraic approach*, London Mathematical Society Lecture Note Series **74**, Cambridge University Press, 1983.
- [24] J.H. van Lint, *Introduction to coding theory*, 3rd ed., Springer, Berlin, 2000.
- [25] J.H. van Lint and R.M. Wilson, *A course in combinatorics*, Cambridge University Press, 1992.
- [26] F.J. MacWilliams and N.J.A. Sloane, *The theory of error-correcting codes*, 2 vols., North-Holland, Amsterdam, 1977.
- [27] Y. Mikami, *The development of mathematics in China and Japan*, 2nd ed., Chelsea, New York, 1974.
- [28] T. Muir, *The theory of determinants in the historical order of development*, reprinted in 2 vols, Dover, New York, 1960.
- [29] M.G. Neubauer and A.J. Ratcliffe, The maximum determinant of  $\pm 1$  matrices, *Linear Algebra Appl.* **257** (1997), 289–306.
- [30] V. Pless, *Introduction to the theory of error-correcting codes*, 3rd ed., Wiley, New York, 1998.
- [31] D. Raghavarao, *Constructions and combinatorial problems in design of experiments*, Wiley, New York, 1971.
- [32] U. Ray, Generalized Kac–Moody algebras and some related topics, *Bull. Amer. Math. Soc. (N.S.)* **38** (2001), 1–42.
- [33] H.J. Ryser, *Combinatorial mathematics*, Mathematical Association of America, 1963.
- [34] J. Seberry and M. Yamada, Hadamard matrices, sequences and block designs, *Contemporary design theory* (ed. J.H. Dinitz and D.R. Stinson), pp. 431–560, Wiley, New York, 1992.
- [35] S. Shahriari, On maximizing  $\det X^t X$ , *Linear and multilinear algebra* **36** (1994), 275–278.
- [36] T.M. Thompson, *From error-correcting codes through sphere packings to simple groups*, Mathematical Association of America, 1983.
- [37] M. Wojtas, On Hadamard’s inequality for the determinants of order non-divisible by 4, *Colloq. Math.* **12** (1964), 73–83.



## VI

### Hensel's $p$ -adic Numbers

The ring  $\mathbb{Z}$  of all integers has a very similar algebraic structure to the ring  $\mathbb{C}[z]$  of all polynomials in one variable with complex coefficients. This similarity extends to their fields of fractions: the field  $\mathbb{Q}$  of rational numbers and the field  $\mathbb{C}(z)$  of rational functions in one variable with complex coefficients. Hensel (1899) had the bold idea of pushing this analogy even further. For any  $\zeta \in \mathbb{C}$ , the ring  $\mathbb{C}[z]$  may be embedded in the ring  $\mathbb{C}_\zeta[[z]]$  of all functions  $f(z) = \sum_{n \geq 0} \alpha_n (z - \zeta)^n$  with complex coefficients  $\alpha_n$  which are holomorphic at  $\zeta$ , and the field  $\mathbb{C}(z)$  may be embedded in the field  $\mathbb{C}_\zeta((z))$  of all functions  $f(z) = \sum_{n \in \mathbb{Z}} \alpha_n (z - \zeta)^n$  with complex coefficients  $\alpha_n$  which are meromorphic at  $\zeta$ , i.e.  $\alpha_n \neq 0$  for at most finitely many  $n < 0$ . Hensel constructed, for each prime  $p$ , a ring  $\mathbb{Z}_p$  of all ' $p$ -adic integers'  $\sum_{n \geq 0} \alpha_n p^n$ , where  $\alpha_n \in \{0, 1, \dots, p-1\}$ , and a field  $\mathbb{Q}_p$  of all ' $p$ -adic numbers'  $\sum_{n \in \mathbb{Z}} \alpha_n p^n$ , where  $\alpha_n \in \{0, 1, \dots, p-1\}$  and  $\alpha_n \neq 0$  for at most finitely many  $n < 0$ . This led him to arithmetic analogues of various analytic results and even to analytic methods of proving them. Hensel's idea of concentrating attention on one prime at a time has proved very fruitful for algebraic number theory. Furthermore, his methods enable the theory of algebraic numbers and the theory of algebraic functions of one variable to be developed completely in parallel.

Hensel simply defined  $p$ -adic integers by their power series expansions. We will adopt a more general approach, due to Kürschák (1913), which is based on absolute values.

#### 1 Valued Fields

Let  $F$  be an arbitrary field. An *absolute value* on  $F$  is a map  $|| : F \rightarrow \mathbb{R}$  with the following properties:

- (V1)  $|0| = 0$ ,  $|a| > 0$  for all  $a \in F$  with  $a \neq 0$ ;
- (V2)  $|ab| = |a||b|$  for all  $a, b \in F$ ;
- (V3)  $|a + b| \leq |a| + |b|$  for all  $a, b \in F$ .

A field with an absolute value will be called simply a *valued field*.

A *non-archimedean absolute value* on  $F$  is a map  $|| : F \rightarrow \mathbb{R}$  with the properties (V1), (V2) and

- (V3)'  $|a + b| \leq \max(|a|, |b|)$  for all  $a, b \in F$ .

A non-archimedean absolute value is indeed an absolute value, since (V1) implies that (V3)' is a strengthening of (V3). An absolute value is said to be *archimedean* if it is not non-archimedean.

The inequality (V3) is usually referred to as the *triangle inequality* and (V3)' as the 'strong triangle', or *ultrametric*, inequality.

If  $F$  is a field with an absolute value  $|\cdot|$ , then the set of real numbers  $|a|$  for all nonzero  $a \in F$  is clearly a subgroup of the multiplicative group of positive real numbers. This subgroup will be called the *value group* of the valued field.

Here are some examples to illustrate these definitions:

(i) An arbitrary field  $F$  has a *trivial* non-archimedean absolute value defined by

$$|0| = 0, \quad |a| = 1 \quad \text{if } a \neq 0.$$

(ii) The ordinary absolute value

$$|a| = a \quad \text{if } a \geq 0, \quad |a| = -a \quad \text{if } a < 0,$$

defines an archimedean absolute value on the field  $\mathbb{Q}$  of rational numbers. We will denote this absolute value by  $|\cdot|_\infty$  to avoid confusion with other absolute values on  $\mathbb{Q}$  which will now be defined.

If  $p$  is a fixed prime, any rational number  $a \neq 0$  can be uniquely expressed in the form  $a = ep^vm/n$ , where  $e = \pm 1$ ,  $v = v_p(a)$  is an integer and  $m, n$  are relatively prime positive integers which are not divisible by  $p$ . It is easily verified that a non-archimedean absolute value is defined on  $\mathbb{Q}$  by putting

$$|0|_p = 0, \quad |a|_p = p^{-v_p(a)} \quad \text{if } a \neq 0.$$

We call this the *p-adic absolute value*.

(iii) Let  $F = K(t)$  be the field of all rational functions in one indeterminate with coefficients from some field  $K$ . Any rational function  $f \neq 0$  can be uniquely expressed in the form  $f = g/h$ , where  $g$  and  $h$  are relatively prime polynomials with coefficients from  $K$  and  $h$  is *monic* (i.e., has leading coefficient 1). If we denote the degrees of  $g$  and  $h$  by  $\partial(g)$  and  $\partial(h)$ , then a non-archimedean absolute value is defined on  $F$  by putting, for a fixed  $q > 1$ ,

$$|0|_\infty = 0, \quad |f|_\infty = q^{\partial(g) - \partial(h)} \quad \text{if } f \neq 0.$$

Other absolute values on  $F$  can be defined in the following way. If  $p \in K[t]$  is a fixed irreducible polynomial, then any rational function  $f \neq 0$  can be uniquely expressed in the form  $f = p^vg/h$ , where  $v = v_p(f)$  is an integer,  $g$  and  $h$  are relatively prime polynomials with coefficients from  $K$  which are not divisible by  $p$ , and  $h$  is monic. It is easily verified that a non-archimedean absolute value is defined on  $F$  by putting, for a fixed  $q > 1$ ,

$$|0|_p = 0, \quad |f|_p = q^{-\partial(p)v_p(f)} \quad \text{if } f \neq 0.$$

(iv) Let  $F = K((t))$  be the field of all formal Laurent series  $f(t) = \sum_{n \in \mathbb{Z}} \alpha_n t^n$  with coefficients  $\alpha_n \in K$  such that  $\alpha_n \neq 0$  for at most finitely many  $n < 0$ . A non-archimedean absolute value is defined on  $F$  by putting, for a fixed  $q > 1$ ,

$$|0| = 0, \quad |f| = q^{-v(f)} \quad \text{if } f \neq 0,$$

where  $v(f)$  is the least integer  $n$  such that  $\alpha_n \neq 0$ .

(v) Let  $F = C_{\mathbb{C}}((z))$  denote the field of all complex-valued functions  $f(z) = \sum_{n \in \mathbb{Z}} \alpha_n (z - \zeta)^n$  which are meromorphic at  $\zeta \in \mathbb{C}$ . Any  $f \in F$  which is not identically zero can be uniquely expressed in the form  $f(z) = (z - \zeta)^v g(z)$ , where  $v = v_{\zeta}(f)$  is an integer,  $g$  is holomorphic at  $\zeta$  and  $g(\zeta) \neq 0$ . A non-archimedean absolute value is defined on  $F$  by putting, for a fixed  $q > 1$ ,

$$|0|_{\zeta} = 0, \quad |f|_{\zeta} = q^{-v_{\zeta}(f)} \quad \text{if } f \neq 0.$$

It should be noted that in examples (iii) and (iv) the restriction of the absolute value to the ground field  $K$  is the trivial absolute value, and the same holds in example (v) for the restriction of the absolute value to  $\mathbb{C}$ . For all the absolute values considered in examples (iii)–(v) the value group is an infinite cyclic group.

We now derive some simple properties common to all absolute values. The notation in the statement of the following lemma is a bit sloppy, since we use the same symbol to denote the unit elements of both  $F$  and  $\mathbb{R}$  (as we have already done for the zero elements).

**Lemma 1** *In any field  $F$  with an absolute value  $|\cdot|$  the following properties hold:*

- (i)  $|1| = 1$ ,  $|-1| = 1$  and, more generally,  $|a| = 1$  for every  $a \in F$  which is a root of unity;
- (ii)  $|-a| = |a|$  for every  $a \in F$ ;
- (iii)  $||a| - |b||_{\infty} \leq |a - b|$  for all  $a, b \in F$ , where  $|\cdot|_{\infty}$  is the ordinary absolute value on  $\mathbb{R}$ ;
- (iv)  $|a^{-1}| = |a|^{-1}$  for every  $a \in F$  with  $a \neq 0$ .

*Proof* By taking  $a = b = 1$  in (V2) and using (V1), we obtain  $|1| = 1$ . If  $a^n = 1$  for some positive integer  $n$ , it now follows from (V2) that  $\alpha = |a|$  satisfies  $\alpha^n = 1$ . Since  $\alpha > 0$ , this implies  $\alpha = 1$ . In particular,  $|-1| = 1$ . Taking  $b = -1$  in (V2), we now obtain (ii).

Replacing  $a$  by  $a - b$  in (V3), we obtain

$$|a| - |b| \leq |a - b|.$$

Since  $a$  and  $b$  may be interchanged, by (ii), this implies (iii). Finally, if we take  $b = a^{-1}$  in (V2) and use (i), we obtain (iv).  $\square$

It follows from Lemma 1(i) that a finite field admits only the trivial absolute value.

We show next how non-archimedean and archimedean absolute values may be distinguished from one another. The notation in the statement of the following proposition is very sloppy, since we use the same symbol to denote both the positive integer  $n$  and the sum  $1 + 1 + \cdots + 1$  ( $n$  summands), although the latter may be 0 if the field has prime characteristic.

**Proposition 2** *Let  $F$  be a field with an absolute value  $|\cdot|$ . Then the following properties are equivalent:*

- (i)  $|2| \leq 1$ ;
- (ii)  $|n| \leq 1$  for every positive integer  $n$ ;
- (iii) the absolute value  $|\cdot|$  is non-archimedean.

*Proof* It is trivial that (iii)  $\Rightarrow$  (i). Suppose now that (i) holds. Then  $|2^k| = |2|^k \leq 1$  for any positive integer  $k$ . An arbitrary positive integer  $n$  can be written to the base 2 in the form

$$n = a_0 + a_1 2 + \cdots + a_g 2^g,$$

where  $a_i \in \{0, 1\}$  for all  $i < g$  and  $a_g = 1$ . Then

$$|n| \leq |a_0| + |a_1| + \cdots + |a_g| \leq g + 1.$$

Now consider the powers  $n^k$ . Since  $n < 2^{g+1}$ , we have  $n^k < 2^{k(g+1)}$  and hence

$$n^k = b_0 + b_1 2 + \cdots + b_h 2^h,$$

where  $b_j \in \{0, 1\}$  for all  $j < h$ ,  $b_h = 1$  and  $h < k(g+1)$ . Thus

$$|n|^k = |n^k| \leq h + 1 \leq k(g+1).$$

Taking  $k$ -th roots and letting  $k \rightarrow \infty$ , we obtain  $|n| \leq 1$ , since  $k^{1/k} = e^{(\log k)/k} \rightarrow 1$  and likewise  $(g+1)^{1/k} = e^{(\log(g+1))/k} \rightarrow 1$ . Thus (i)  $\Rightarrow$  (ii).

Suppose next that (ii) holds. Then, since the binomial coefficients are positive integers,

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

where  $\rho = \max(|x|, |y|)$ . Taking  $n$ -th roots and letting  $n \rightarrow \infty$ , we obtain  $|x + y| \leq \rho$ . Thus (ii)  $\Rightarrow$  (iii).  $\square$

It follows from Proposition 2 that for an archimedean absolute value the sequence  $(|n|)$  is unbounded, since  $|2^k| \rightarrow \infty$  as  $k \rightarrow \infty$ . Consequently, for any  $a, b \in F$  with  $a \neq 0$ , there is a positive integer  $n$  such that  $|na| > |b|$ . The name ‘archimedean’ is used because of the analogy with the archimedean axiom of geometry. It follows also from Proposition 2 that any absolute value on a field of prime characteristic is non-archimedean, since there are only finitely many distinct values of  $|n|$ .

## 2 Equivalence

If  $\lambda, \mu, \alpha$  are positive real numbers with  $\alpha < 1$ , then

$$\left(\frac{\lambda}{\lambda + \mu}\right)^\alpha + \left(\frac{\mu}{\lambda + \mu}\right)^\alpha > \frac{\lambda}{\lambda + \mu} + \frac{\mu}{\lambda + \mu} = 1$$

and hence

$$\lambda^\alpha + \mu^\alpha > (\lambda + \mu)^\alpha.$$

It follows that if  $||$  is an absolute value on a field  $F$  and if  $0 < \alpha < 1$ , then  $||^\alpha$  is also an absolute value, since

$$|a + b|^\alpha \leq (|a| + |b|)^\alpha \leq |a|^\alpha + |b|^\alpha.$$

Actually, if  $||$  is a *non-archimedean* absolute value on a field  $F$ , then it follows directly from the definition that, for any  $\alpha > 0$ ,  $||^\alpha$  is also a non-archimedean absolute value on  $F$ . However, if  $||$  is an *archimedean* absolute value on  $F$  then, for all large  $\alpha > 0$ ,  $||^\alpha$  is not an absolute value on  $F$ . For  $|2| > 1$  and hence, if  $\alpha > \log 2 / \log |2|$ ,

$$|1 + 1|^\alpha > 2 = |1|^\alpha + |1|^\alpha.$$

**Proposition 3** Let  $||_1$  and  $||_2$  be absolute values on a field  $F$  such that  $|a|_2 < 1$  for any  $a \in F$  with  $|a|_1 < 1$ . If  $||_1$  is nontrivial, then there exists a real number  $\rho > 0$  such that

$$|a|_2 = |a|_1^\rho \quad \text{for every } a \in F.$$

*Proof* By taking inverses we see that also  $|a|_2 > 1$  for any  $a \in F$  with  $|a|_1 > 1$ . Choose  $b \in F$  with  $|b|_1 > 1$ . For any nonzero  $a \in F$  we have  $|a|_1 = |b|_1^\gamma$ , where

$$\gamma = \log |a|_1 / \log |b|_1.$$

Let  $m, n$  be integers with  $n > 0$  such that  $m/n > \gamma$ . Then  $|a|_1^m < |b|_1^n$  and hence  $|a^n/b^m|_1 < 1$ . Therefore also  $|a^n/b^m|_2 < 1$  and by reversing the argument we obtain

$$m/n > \log |a|_2 / \log |b|_2.$$

Similarly if  $m', n'$  are integers with  $n' > 0$  such that  $m'/n' < \gamma$ , then

$$m'/n' < \log |a|_2 / \log |b|_2.$$

It follows that

$$\log |a|_2 / \log |b|_2 = \gamma = \log |a|_1 / \log |b|_1.$$

Thus if we put  $\rho = \log |b|_2 / \log |b|_1$ , then  $\rho > 0$  and  $|a|_2 = |a|_1^\rho$ . This holds trivially also for  $a = 0$ .  $\square$

Two absolute values,  $|\cdot|_1$  and  $|\cdot|_2$ , on a field  $F$  are said to be *equivalent* when, for any  $a \in F$ ,

$$|a|_1 < 1 \quad \text{if and only if} \quad |a|_2 < 1.$$

This implies that  $|a|_1 > 1$  if and only if  $|a|_2 > 1$  and hence also that  $|a|_1 = 1$  if and only if  $|a|_2 = 1$ . Thus if one absolute value is trivial, so also is the other. It now follows from Proposition 3 that *two absolute values,  $|\cdot|_1$  and  $|\cdot|_2$ , on a field  $F$  are equivalent if and only if there exists a real number  $\rho > 0$  such that  $|a|_2 = |a|_1^\rho$  for every  $a \in F$ .*

We have seen that the field  $\mathbb{Q}$  of rational numbers admits the  $p$ -adic absolute values  $|\cdot|_p$  in addition to the ordinary absolute value  $|\cdot|_\infty$ . These absolute values are all inequivalent since, if  $p$  and  $q$  are distinct primes,

$$|p|_p < 1, \quad |p|_q = 1, \quad |p|_\infty = p > 1.$$

It was first shown by Ostrowski (1918) that these are essentially the only absolute values on  $\mathbb{Q}$ :

**Proposition 4** *Every nontrivial absolute value  $|\cdot|$  of the rational field  $\mathbb{Q}$  is equivalent either to the ordinary absolute value  $|\cdot|_\infty$  or to a  $p$ -adic absolute value  $|\cdot|_p$  for some prime  $p$ .*

*Proof* Let  $b, c$  be integers  $> 1$ . By writing  $c$  to the base  $b$ , we obtain

$$c = c_m b^m + c_{m-1} b^{m-1} + \cdots + c_0,$$

where  $0 \leq c_j < b$  ( $j = 0, \dots, m$ ) and  $c_m \neq 0$ . Then  $m \leq \log c / \log b$ , since  $c_m \geq 1$ . If we put  $\mu = \max_{1 \leq d < b} |d|$ , it follows from the triangle inequality that

$$|c| \leq \mu(1 + \log c / \log b) \{\max(1, |b|)\}^{\log c / \log b}.$$

Taking  $c = a^n$  we obtain, for any  $a > 1$ ,

$$|a| \leq \mu^{1/n} (1 + n \log a / \log b)^{1/n} \{\max(1, |b|)\}^{\log a / \log b}$$

and hence, letting  $n \rightarrow \infty$ ,

$$|a| \leq \{\max(1, |b|)\}^{\log a / \log b}.$$

Suppose first that  $|a| > 1$  for some  $a > 1$ . It follows that  $|b| > 1$  for every  $b > 1$  and

$$|b|^{1/\log b} \geq |a|^{1/\log a}.$$

In fact, since  $a$  and  $b$  may now be interchanged,

$$|b|^{1/\log b} = |a|^{1/\log a}.$$

Thus  $\rho = \log |a| / \log a$  is a positive real number independent of  $a > 1$  and  $|a| = a^\rho$ . It follows that  $|a| = |a|_\infty^\rho$  for every rational number  $a$ . Thus the absolute value is equivalent to the ordinary absolute value.

Suppose next that  $|a| \leq 1$  for every  $a > 1$  and so for every  $a \in \mathbb{Z}$ . Since the absolute value on  $\mathbb{Q}$  is nontrivial, we must have  $|a| < 1$  for some integer  $a \neq 0$ . The set  $M$  of all  $a \in \mathbb{Z}$  such that  $|a| < 1$  is a proper ideal in  $\mathbb{Z}$  and hence is generated by an integer  $p > 1$ . We will show that  $p$  must be a prime. Suppose  $p = bc$ , where  $b$  and  $c$  are positive integers. Since  $|b||c| = |p| < 1$ , we may assume without loss of generality that  $|b| < 1$ . Then  $b \in M$  and thus  $b = pd$  for some  $d \in \mathbb{Z}$ . Hence  $cd = 1$  and so  $c = 1$ . Thus  $p$  has no nontrivial factorization.

Every rational number  $a \neq 0$  can be expressed in the form  $a = p^v b/c$ , where  $v$  is an integer and  $b, c$  are integers not divisible by  $p$ . Hence  $|b| = |c| = 1$  and  $|a| = |p|^v$ . We can write  $|p| = p^{-\rho}$ , for some real number  $\rho > 0$ . Then  $|a| = p^{-v\rho} = |a|_p^\rho$ , and thus the absolute value is equivalent to the  $p$ -adic absolute value.  $\square$

Similarly, the absolute values on the field  $F = K(t)$  considered in example (iii) of §1 are all inequivalent and it may be shown that any nontrivial absolute value on  $F$  whose restriction to  $K$  is trivial is equivalent to one of these absolute values.

In example (ii) of §1 we have made a specific choice in each class of equivalent absolute values. The choice which has been made ensures the validity of the *product formula*: for any nonzero  $a \in \mathbb{Q}$ ,

$$|a|_\infty \prod_p |a|_p = 1,$$

where  $|a|_p \neq 1$  for at most finitely many  $p$ .

Similarly, in example (iii) of §1 the absolute values have been chosen so that, for any nonzero  $f \in K(t)$ ,  $|f|_\infty \prod_p |f|_p = 1$ , where  $|f|_p \neq 1$  for at most finitely many  $p$ .

The following *approximation theorem*, due to Artin and Whaples (1945), treats several absolute values simultaneously. For  $p$ -adic absolute values of the rational field  $\mathbb{Q}$  the result also follows from the Chinese remainder theorem (Corollary II.38).

**Proposition 5** *Let  $| \cdot |_1, \dots, | \cdot |_m$  be nontrivial pairwise inequivalent absolute values of an arbitrary field  $F$  and let  $x_1, \dots, x_m$  be any elements of  $F$ . Then for each real  $\varepsilon > 0$  there exists an  $x \in F$  such that*

$$|x - x_k|_k < \varepsilon \quad \text{for } 1 \leq k \leq m.$$

*Proof* During the proof we will more than once use the fact that if  $f_n(x) = x^n(1 + x^n)^{-1}$ , then  $|f_n(a)| \rightarrow 0$  or  $1$  as  $n \rightarrow \infty$  according as  $|a| < 1$  or  $|a| > 1$ .

We show first that *there exists an  $a \in F$  such that*

$$|a|_1 > 1, \quad |a|_k < 1 \quad \text{for } 2 \leq k \leq m.$$

Since  $| \cdot |_1$  and  $| \cdot |_2$  are nontrivial and inequivalent, there exist  $b, c \in F$  such that

$$\begin{aligned} |b|_1 < 1, \quad |b|_2 &\geq 1, \\ |c|_1 &\geq 1, \quad |c|_2 < 1. \end{aligned}$$

If we put  $a = b^{-1}c$ , then  $|a|_1 > 1$ ,  $|a|_2 < 1$ . This proves the assertion for  $m = 2$ . We now assume  $m > 2$  and use induction. Then there exist  $b, c \in F$  such that

$$\begin{aligned} |b|_1 > 1, \quad |b|_k < 1 \quad \text{for } 1 < k < m, \\ |c|_1 > 1, \quad |c|_m < 1. \end{aligned}$$

If  $|b|_m < 1$  we can take  $a = b$ . If  $|b|_m = 1$  we can take  $a = b^n c$  for sufficiently large  $n$ . If  $|b|_m > 1$  we can take  $a = f_n(b)c$  for sufficiently large  $n$ .

Thus for each  $i \in \{1, \dots, m\}$  we can choose  $a_i \in F$  so that

$$|a_i|_i > 1, \quad |a_i|_k < 1 \quad \text{for all } k \neq i.$$

Then

$$x = x_1 f_n(a_1) + \dots + x_m f_n(a_m)$$

satisfies the requirements of the proposition for sufficiently large  $n$ .  $\square$

It follows from Proposition 5, that if  $|\cdot|_1, \dots, |\cdot|_m$  are nontrivial pairwise inequivalent absolute values of a field  $F$ , then there exists an  $a \in F$  such that  $|a|_k > 1$  ( $k = 1, \dots, m$ ). Consequently the absolute values are *multiplicatively independent*, i.e. if  $\rho_1, \dots, \rho_m$  are nonnegative real numbers, not all zero, then for some nonzero  $a \in F$

$$|a|_1^{\rho_1} \cdots |a|_m^{\rho_m} \neq 1.$$

### 3 Completions

Any field  $F$  with an absolute value  $|\cdot|$  has the structure of a metric space, with the metric

$$d(a, b) = |a - b|,$$

and thus has an associated topology. Since  $|a| < 1$  if and only if  $a^n \rightarrow 0$  as  $n \rightarrow \infty$ , it follows that two absolute values are equivalent if and only if the induced topologies are the same.

When we use topological concepts in connection with valued fields we will always refer to the topology induced by the metric space structure. In this sense addition and multiplication are continuous operations, since

$$\begin{aligned} |(a + b) - (a_0 + b_0)| &\leq |a - a_0| + |b - b_0|, \\ |ab - a_0b_0| &\leq |a - a_0||b| + |a_0||b - b_0|. \end{aligned}$$

Inversion is also continuous at any point  $a_0 \neq 0$ , since if  $|a - a_0| < |a_0|/2$  then  $|a_0| < 2|a|$  and

$$|a^{-1} - a_0^{-1}| = |a - a_0||a|^{-1}|a_0|^{-1} < 2|a_0|^{-2}|a - a_0|.$$

Thus a valued field is a *topological field*.

It will now be shown that the procedure by which Cantor extended the field of rational numbers to the field of real numbers can be generalized to any valued field.

Let  $F$  be a field with an absolute value  $|\cdot|$ . A sequence  $(a_n)$  of elements of  $F$  is said to *converge* to an element  $a$  of  $F$ , and  $a$  is said to be the *limit* of the sequence  $(a_n)$ , if for each real  $\varepsilon > 0$  there is a corresponding positive integer  $N = N(\varepsilon)$  such that

$$|a_n - a| < \varepsilon \quad \text{for all } n \geq N.$$

It is easily seen that the limit of a convergent sequence is uniquely determined.

A sequence  $(a_n)$  of elements of  $F$  is said to be a *fundamental sequence* if for each  $\varepsilon > 0$  there is a corresponding positive integer  $N = N(\varepsilon)$  such that

$$|a_m - a_n| < \varepsilon \quad \text{for all } m, n \geq N.$$

Any convergent sequence is a fundamental sequence, since

$$|a_m - a_n| \leq |a_m - a| + |a_n - a|,$$

but the converse need not hold. However, any fundamental sequence is bounded since, if  $m = N(1)$ , then for  $n \geq m$  we have

$$|a_n| \leq |a_m - a_n| + |a_m| < 1 + |a_m|.$$

Thus  $|a_n| \leq \mu$  for all  $n$ , where  $\mu = \max\{|a_1|, \dots, |a_{m-1}|, 1 + |a_m|\}$ .

The preceding definitions are specializations of the definitions for an arbitrary metric space (cf. Chapter I, §4). We now take advantage of the algebraic structure of  $F$ . Let  $A = (a_n)$  and  $B = (b_n)$  be two fundamental sequences. We write  $A = B$  if  $a_n = b_n$  for all  $n$ , and we define the sum and product of  $A$  and  $B$  to be the sequences

$$A + B = (a_n + b_n), \quad AB = (a_n b_n).$$

These are again fundamental sequences. For we can choose  $\mu \geq 1$  so that  $|a_n| \leq \mu$ ,  $|b_n| \leq \mu$  for all  $n$  and then choose a positive integer  $N$  so that

$$|a_m - a_n| < \varepsilon/2\mu, \quad |b_m - b_n| < \varepsilon/2\mu \quad \text{for all } m, n \geq N.$$

It follows that, for all  $m, n \geq N$ ,

$$|(a_m + b_m) - (a_n + b_n)| \leq |a_m - a_n| + |b_m - b_n| < \varepsilon/2\mu + \varepsilon/2\mu \leq \varepsilon,$$

and similarly

$$|a_m b_m - a_n b_n| \leq |a_m - a_n| |b_m| + |a_n| |b_m - b_n| < (\varepsilon/2\mu)\mu + (\varepsilon/2\mu)\mu = \varepsilon.$$

It is easily seen that the set  $\mathcal{F}$  of all fundamental sequences is a commutative ring with respect to these operations. The subset of all constant sequences  $(a)$ , i.e.  $a_n = a$  for all  $n$ , forms a field isomorphic to  $F$ . Thus we may regard  $F$  as embedded in  $\mathcal{F}$ .

Let  $\mathcal{N}$  denote the subset of  $\mathcal{F}$  consisting of all sequences  $(a_n)$  which converge to 0. Evidently  $\mathcal{N}$  is a subring of  $\mathcal{F}$  and actually an ideal, since any fundamental sequence is bounded. We will show that  $\mathcal{N}$  is even a maximal ideal.

Let  $(a_n)$  be a fundamental sequence which is not in  $\mathcal{N}$ . Then there exists  $\mu > 0$  such that  $|a_v| \geq \mu$  for infinitely many  $v$ . Since  $|a_m - a_n| < \mu/2$  for all  $m, n \geq N$ , it follows that  $|a_n| > \mu/2$  for all  $n \geq N$ . Put  $b_n = a_n^{-1}$  if  $a_n \neq 0$ ,  $b_n = 0$  if  $a_n = 0$ . Then  $(b_n)$  is a fundamental sequence since, for  $m, n \geq N$ ,

$$|b_m - b_n| = |(a_n - a_m)/a_m a_n| \leq 4\mu^{-2}|a_n - a_m|.$$

Since  $(1) - (b_n a_n) \in \mathcal{N}$ , the ideal generated by  $(a_n)$  and  $\mathcal{N}$  contains the constant sequence  $(1)$  and hence every sequence in  $\mathcal{F}$ . Since this holds for each sequence  $(a_n) \in \mathcal{F} \setminus \mathcal{N}$ , the ideal  $\mathcal{N}$  is maximal.

Consequently (see Chapter I, §8) the quotient  $\bar{F} = \mathcal{F}/\mathcal{N}$  is a field. Since  $(0)$  is the only constant sequence in  $\mathcal{N}$ , by mapping each constant sequence into the coset of  $\mathcal{N}$  which contains it we obtain a field in  $\bar{F}$  isomorphic to  $F$ . Thus we may regard  $F$  as embedded in  $\bar{F}$ .

It follows from Lemma 1(iii), and from the completeness of the field of real numbers, that  $|A| = \lim_{n \rightarrow \infty} |a_n|$  exists for any fundamental sequence  $A = (a_n)$ . Moreover,

$$|A| \geq 0, \quad |AB| = |A||B|, \quad |A + B| \leq |A| + |B|.$$

Furthermore  $|A| = 0$  if and only if  $A \in \mathcal{N}$ . It follows that  $|B| = |C|$  if  $B - C \in \mathcal{N}$ , since

$$|B| \leq |B - C| + |C| = |C| \leq |C - B| + |B| = |B|.$$

Thus we may consider  $||$  as defined on  $\bar{F} = \mathcal{F}/\mathcal{N}$ , and it is then an absolute value on the field  $\bar{F}$  which coincides with the original absolute value when restricted to the field  $F$ .

If  $A = (a_n)$  is a fundamental sequence, and if  $A_m$  is the constant sequence  $(a_m)$ , then  $|A - A_m|$  can be made arbitrarily small by taking  $m$  sufficiently large. It follows that  $F$  is *dense* in  $\bar{F}$ , i.e. for any  $\alpha \in \bar{F}$  and any  $\varepsilon > 0$  there exists  $a \in F$  such that  $|\alpha - a| < \varepsilon$ .

We show finally that  $\bar{F}$  is *complete* as a metric space, i.e. every fundamental sequence of elements of  $\bar{F}$  converges to an element of  $\bar{F}$ . For let  $(\alpha_n)$  be a fundamental sequence in  $\bar{F}$ . Since  $F$  is dense in  $\bar{F}$ , for each  $n$  we can choose  $a_n \in F$  so that  $|\alpha_n - a_n| < 1/n$ . Since

$$|a_m - a_n| \leq |a_m - \alpha_m| + |\alpha_m - \alpha_n| + |\alpha_n - a_n|,$$

it follows that  $(a_n)$  is also a fundamental sequence. Thus there exists  $\alpha \in \bar{F}$  such that  $\lim_{n \rightarrow \infty} |a_n - \alpha| = 0$ . Since

$$|\alpha_n - \alpha| \leq |\alpha_n - a_n| + |a_n - \alpha|,$$

we have also  $\lim_{n \rightarrow \infty} |\alpha_n - \alpha| = 0$ . Thus the sequence  $(\alpha_n)$  converges to  $\alpha$ .

Summing up, we have proved

**Proposition 6** *If  $F$  is a field with an absolute value  $||$ , then there exists a field  $\bar{F}$  containing  $F$ , with an absolute value  $||$  extending that of  $F$ , such that  $\bar{F}$  is complete and  $F$  is dense in  $\bar{F}$ .*

It is easily seen that  $\bar{F}$  is uniquely determined, up to an isomorphism which preserves the absolute value. The field  $\bar{F}$  is called the *completion* of the valued field  $F$ . The density of  $F$  in  $\bar{F}$  implies that the absolute value on the completion  $\bar{F}$  is non-archimedean or archimedean according as the absolute value on  $F$  is non-archimedean or archimedean.

It is easy to see that in example (iv) of §1 the valued field  $F = K((t))$  of all formal Laurent series is complete, i.e. it is its own completion. For let  $\{f^{(k)}\}$  be a fundamental sequence in  $F$ . Given any positive integer  $N$ , there is a positive integer  $M = M(N)$  such that  $|f^{(k)} - f^{(j)}| < q^{-N}$  for  $j, k \geq M$ . Thus we can write

$$f^{(k)}(t) = \sum_{n \leq N} \alpha_n t^n + \sum_{n > N} \alpha_n^{(k)} t^n \quad \text{for all } k \geq M.$$

If  $f(t) = \sum_{n \in \mathbb{Z}} \alpha_n t^n$ , then  $\lim_{k \rightarrow \infty} |f^{(k)} - f| = 0$ .

On the other hand, given any  $f(t) = \sum_{n \in \mathbb{Z}} \alpha_n t^n \in K((t))$ , we have  $|f^{(k)} - f| \rightarrow 0$  as  $k \rightarrow \infty$ , where  $f^{(k)}(t) = \sum_{n \leq k} \alpha_n t^n \in K(t)$ . It follows that  $K((t))$  is the completion of the field  $K(t)$  of rational functions considered in example (iii) of §1, with the absolute value  $|\cdot|_t$  corresponding to the irreducible polynomial  $p(t) = t$  (for which  $\partial(p) = 1$ ).

The completion of the rational field  $\mathbb{Q}$  with respect to the  $p$ -adic absolute value  $|\cdot|_p$  will be denoted by  $\mathbb{Q}_p$ , and the elements of  $\mathbb{Q}_p$  will be called *p-adic numbers*.

The completion of the rational field  $\mathbb{Q}$  with respect to the ordinary absolute value  $|\cdot|_\infty$  is of course the real field  $\mathbb{R}$ . In §6 we will show that the only fields with a complete archimedean absolute value are the real field  $\mathbb{R}$  and the complex field  $\mathbb{C}$ , and the absolute value has the form  $|\cdot|_\infty^\rho$  for some  $\rho > 0$ . In fact  $\rho \leq 1$ , since  $2^\rho \leq 1^\rho + 1^\rho = 2$ . Thus an arbitrary archimedean valued field is equivalent to a subfield of  $\mathbb{C}$  with the usual absolute value. (Hence, for a field with an archimedean absolute value  $|\cdot|$ ,  $|n| > 1$  for every integer  $n > 1$  and  $|n| \rightarrow \infty$  as  $n \rightarrow \infty$ .) Since this case may be considered well-known, we will in the following devote our attention primarily to the peculiarities of non-archimedean valued fields.

We will later be concerned with extending an absolute value on a field  $F$  to a field  $E$  which is a finite extension of  $F$ . Since all that matters for some purposes is that  $E$  is a vector space over  $F$ , it is useful to introduce the following definition.

Let  $F$  be a field with an absolute value  $|\cdot|$  and let  $E$  be a vector space over  $F$ . A *norm* on  $E$  is a map  $\|\cdot\| : E \rightarrow \mathbb{R}$  with the following properties:

- (i)  $\|a\| > 0$  for every  $a \in E$  with  $a \neq 0$ ;
- (ii)  $\|\alpha a\| = |\alpha| \|a\|$  for all  $\alpha \in F$  and  $a \in E$ ;
- (iii)  $\|a + b\| \leq \|a\| + \|b\|$  for all  $a, b \in E$ .

It follows from (ii) that  $\|0\| = 0$ . We will require only one result about normed vector spaces:

**Lemma 7** *Let  $F$  be a complete valued field and let  $E$  be a finite-dimensional vector space over  $F$ . If  $\|\cdot\|_1$  and  $\|\cdot\|_2$  are both norms on  $E$ , then there exist positive constants  $\sigma, \mu$  such that*

$$\sigma \|a\|_1 \leq \|a\|_2 \leq \mu \|a\|_1 \quad \text{for every } a \in E.$$

*Proof* Let  $e_1, \dots, e_n$  be a basis for the vector space  $E$ . Then any  $a \in E$  can be uniquely represented in the form

$$a = \alpha_1 e_1 + \dots + \alpha_n e_n,$$

where  $\alpha_1, \dots, \alpha_n \in F$ . It is easily seen that

$$\|a\|_0 = \max_{1 \leq i \leq n} |\alpha_i|$$

is a norm on  $E$ , and it is sufficient to prove the proposition for  $\|\cdot\|_2 = \|\cdot\|_0$ . Since

$$\|a\|_1 \leq \|a\|_0(\|e_1\|_1 + \dots + \|e_n\|_1),$$

we can take  $\sigma = (\|e_1\|_1 + \dots + \|e_n\|_1)^{-1}$ . To establish the existence of  $\mu$  we assume  $n > 1$  and use induction, since the result is obviously true for  $n = 1$ .

Assume, contrary to the assertion, that there exists a sequence  $a^{(k)} \in E$  such that

$$\|a^{(k)}\|_1 < \varepsilon_k \|a^{(k)}\|_0,$$

where  $\varepsilon_k > 0$  and  $\varepsilon_k \rightarrow 0$  as  $k \rightarrow \infty$ . We may suppose, without loss of generality, that

$$|\alpha_n^{(k)}| = \|a^{(k)}\|_0$$

and also, by replacing  $a^{(k)}$  by  $(\alpha_n^{(k)})^{-1} a^{(k)}$ , that  $\alpha_n^{(k)} = 1$ . Thus  $a^{(k)} = b^{(k)} + e_n$ , where

$$b^{(k)} = \alpha_1^{(k)} e_1 + \dots + \alpha_{n-1}^{(k)} e_{n-1},$$

and  $\|a^{(k)}\|_1 \rightarrow 0$  as  $k \rightarrow \infty$ . The sequences  $\alpha_i^{(k)}$  ( $i = 1, \dots, n-1$ ) are fundamental sequences in  $F$ , since

$$\|b^{(j)} - b^{(k)}\|_1 \leq \|b^{(j)} + e_n\|_1 + \|b^{(k)} + e_n\|_1 = \|a^{(j)}\|_1 + \|a^{(k)}\|_1$$

and, by the induction hypothesis,

$$|\alpha_i^{(j)} - \alpha_i^{(k)}| \leq \mu_{n-1} \|b^{(j)} - b^{(k)}\|_1 \quad (i = 1, \dots, n-1).$$

Hence, since  $F$  is complete, there exist  $\alpha_i \in F$  such that  $|\alpha_i^{(k)} - \alpha_i| \rightarrow 0$  ( $i = 1, \dots, n-1$ ). Put

$$b = \alpha_1 e_1 + \dots + \alpha_{n-1} e_{n-1}.$$

Since  $\|b^{(k)} - b\|_1 \leq \sigma_{n-1}^{-1} \|b^{(k)} - b\|_0$ , it follows that  $\|b^{(k)} - b\|_1 \rightarrow 0$ . But if  $a = b + e_n$ , then

$$\|a\|_1 \leq \|a - a^{(k)}\|_1 + \|a^{(k)}\|_1 = \|b - b^{(k)}\|_1 + \|a^{(k)}\|_1.$$

Letting  $k \rightarrow \infty$ , we obtain  $a = 0$ , which contradicts the definition of  $a$ .  $\square$

#### 4 Non-Archimedean Valued Fields

Throughout this section we denote by  $F$  a field with a non-archimedean absolute value  $|\cdot|$ . A basic property of such fields is the following simple lemma. It may be interpreted as saying that in ultrametric geometry every triangle is isosceles.

**Lemma 8** *If  $a, b \in F$  and  $|a| < |b|$ , then  $|a + b| = |b|$ .*

*Proof* We certainly have

$$|a + b| \leq \max\{|a|, |b|\} = |b|.$$

On the other hand, since  $b = (a + b) - a$ , we have

$$|b| \leq \max\{|a + b|, |-a|\}$$

and, since  $|-a| = |a| < |b|$ , this implies  $|b| \leq |a + b|$ .  $\square$

It may be noted that if  $a \neq 0$  and  $b = -a$ , then  $|a| = |b|$  and  $|a + b| < |b|$ . From Lemma 8 it follows by induction that if  $a_1, \dots, a_n \in F$  and  $|a_k| < |a_1|$  for  $1 < k \leq n$ , then

$$|a_1 + \dots + a_n| = |a_1|.$$

As an application we show that if a field  $E$  is a finite extension of a field  $F$ , then the trivial absolute value on  $E$  is the only extension to  $E$  of the trivial absolute value on  $F$ . By Proposition 2, any extension to  $E$  of the trivial absolute value on  $F$  must be non-archimedean. Suppose  $\alpha \in E$  and  $|\alpha| > 1$ . Then  $\alpha$  satisfies a polynomial equation

$$\alpha^n + c_{n-1}\alpha^{n-1} + \dots + c_0 = 0$$

with coefficients  $c_k \in F$ . Since  $|c_k| = 0$  or  $1$  and since  $|\alpha^k| < |\alpha^n|$  if  $k < n$ , we obtain the contradiction  $|\alpha^n| = |\alpha^n + c_{n-1}\alpha^{n-1} + \dots + c_0| = 0$ .

As another application we prove

**Proposition 9** *If a field  $F$  has a non-archimedean absolute value  $|\cdot|$ , then the valuation on  $F$  can be extended to the polynomial ring  $F[t]$  by defining the absolute value of  $f(t) = a_0 + a_1t + \dots + a_nt^n$  to be  $|f| = \max\{|a_0|, \dots, |a_n|\}$ .*

*Proof* We need only show that  $|fg| = |f||g|$ , since it is evident that  $|f| = 0$  if and only if  $f = 0$  and that  $|f + g| \leq |f| + |g|$ . Let  $g(t) = b_0 + b_1t + \dots + b_mt^m$ . Then  $f(t)g(t) = c_0 + c_1t + \dots + c_lt^l$ , where

$$c_i = a_0b_i + a_1b_{i-1} + \dots + a_ib_0.$$

If  $r$  is the least integer such that  $|a_r| = |f|$  and  $s$  the least integer such that  $|b_s| = |g|$ , then  $a_rb_s$  has strictly greatest absolute value among all products  $a_jb_k$  with  $j + k = r + s$ . Hence  $|c_{r+s}| = |a_r||b_s|$  and  $|fg| \geq |f||g|$ . On the other hand,

$$|fg| = \max_i |c_i| \leq \max_{j,k} |a_j||b_k| = |f||g|.$$

Consequently  $|fg| = |f||g|$ . Clearly also  $|f| = |a|$  if  $f = a \in F$ . (The absolute value on  $F$  can be further extended to the field  $F(t)$  of rational functions by defining  $|f(t)/g(t)|$  to be  $|f|/|g|$ .)  $\square$

It also follows at once from Lemma 8 that if a sequence  $(a_n)$  of elements of  $F$  converges to a limit  $a \neq 0$ , then  $|a_n| = |a|$  for all large  $n$ . Hence the value group of the field  $F$  is the same as the value group of its completion  $\bar{F}$ . The next lemma has an especially appealing corollary.

**Lemma 10** *Let  $F$  be a field with a non-archimedean absolute value  $||$ . Then a sequence  $(a_n)$  of elements of  $F$  is a fundamental sequence if and only if  $\lim_{n \rightarrow \infty} |a_{n+1} - a_n| = 0$ .*

*Proof* If  $|a_{n+1} - a_n| \rightarrow 0$ , then for each  $\varepsilon > 0$  there is a corresponding positive integer  $N = N(\varepsilon)$  such that

$$|a_{n+1} - a_n| < \varepsilon \quad \text{for } n \geq N.$$

For any integer  $k > 1$ ,

$$a_{n+k} - a_n = (a_{n+1} - a_n) + (a_{n+2} - a_{n+1}) + \cdots + (a_{n+k} - a_{n+k-1})$$

and hence

$$|a_{n+k} - a_n| \leq \max\{|a_{n+1} - a_n|, |a_{n+2} - a_{n+1}|, \dots, |a_{n+k} - a_{n+k-1}|\} < \varepsilon \text{ for } n \geq N.$$

Thus  $(a_n)$  is a fundamental sequence. The converse follows at once from the definition of a fundamental sequence.  $\square$

**Corollary 11** *In a field  $F$  with a complete non-archimedean absolute value  $||$ , an infinite series  $\sum_{n=1}^{\infty} a_n$  of elements of  $F$  is convergent if and only if  $|a_n| \rightarrow 0$ .*

Let  $F$  be a field with a nontrivial non-archimedean absolute value  $||$  and put

$$R = \{a \in F : |a| \leq 1\},$$

$$M = \{a \in F : |a| < 1\},$$

$$U = \{a \in F : |a| = 1\}.$$

Then  $R$  is the union of the disjoint nonempty subsets  $M$  and  $U$ . It follows from the definition of a non-archimedean absolute value that  $R$  is a (commutative) ring containing the unit element of  $F$  and that, for any nonzero  $a \in F$ , either  $a \in R$  or  $a^{-1} \in R$  (or both). Moreover  $M$  is an ideal of  $R$  and  $U$  is a multiplicative group, consisting of all  $a \in R$  such that also  $a^{-1} \in R$ . Thus a proper ideal of  $R$  cannot contain an element of  $U$  and hence  $M$  is the unique maximal ideal of  $R$ . Consequently (see again Chapter I, §8) the quotient  $R/M$  is a field.

We call  $R$  the *valuation ring*,  $M$  the *valuation ideal*, and  $R/M$  the *residue field* of the valued field  $F$ .

We draw attention to the fact that the ‘closed unit ball’  $R$  is both open and closed in the topology induced by the absolute value. For if  $a \in R$  and  $|b - a| < 1$ , then also  $b \in R$ . Furthermore, if  $a_n \in R$  and  $a_n \rightarrow a$  then  $a \in R$ , since  $|a_n| = |a|$  for all large  $n$ . Similarly, the ‘open unit ball’  $M$  is also both open and closed.

In particular, let  $F = \mathbb{Q}$  be the field of rational numbers and  $|| = ||_p$  the  $p$ -adic absolute value. In this case the valuation ring  $R = R_p$  is the set of all rational numbers

$m/n$ , where  $m$  and  $n$  are relatively prime integers,  $n > 0$  and  $p$  does not divide  $n$ . The valuation ideal is  $M = pR_p$  and the residue field  $\mathbb{F}_p = R_p/pR_p$  is the finite field with  $p$  elements.

As another example, let  $F = K(t)$  be the field of rational functions with coefficients from an arbitrary field  $K$  and let  $|| = ||_t$  be the absolute value considered in example (iii) of §1 for the irreducible polynomial  $p(t) = t$ . In this case the valuation ring  $R$  is the set of all rational functions  $f = g/h$ , where  $g$  and  $h$  are relatively prime polynomials and  $h$  has nonzero constant term. The valuation ideal is  $M = tR$  and the residue field  $R/M$  is isomorphic to  $K$ , since  $f(t) \equiv f(0) \pmod{M}$  (i.e.,  $f(t) - f(0) \in M$ ).

Let  $\bar{F}$  be the completion of  $F$ . If  $\bar{R}$  and  $\bar{M}$  are the valuation ring and valuation ideal of  $\bar{F}$ , then evidently

$$R = \bar{R} \cap F, \quad M = \bar{M} \cap F.$$

Moreover  $R$  is dense in  $\bar{R}$  since, if  $0 < \varepsilon \leq 1$ , for any  $\alpha \in \bar{R}$  there exists  $a \in F$  such that  $|\alpha - a| < \varepsilon$  and then  $a \in R$  (and  $\alpha - a \in \bar{M}$ ). Furthermore the residue fields  $R/M$  and  $\bar{R}/\bar{M}$  are isomorphic. For the map  $a + M \rightarrow a + \bar{M}$  ( $a \in R$ ) is an isomorphism of  $R/M$  onto a subfield of  $\bar{R}/\bar{M}$  and this subfield is not proper (by the preceding bracketed remark).

The valuation ring of the field  $\mathbb{Q}_p$  of  $p$ -adic numbers will be denoted by  $\mathbb{Z}_p$  and its elements will be called *p-adic integers*. The ring  $\mathbb{Z}$  of ordinary integers is dense in  $\mathbb{Z}_p$ , and the residue field of  $\mathbb{Q}_p$  is the finite field  $\mathbb{F}_p$  with  $p$  elements, since this is the residue field of  $\mathbb{Q}$ .

Similarly, the valuation ring of the field  $K((t))$  of all formal Laurent series is the ring  $K[[t]]$  of all formal power series  $\sum_{n \geq 0} a_n t^n$ . The polynomial ring  $K[t]$  is dense in  $K[[t]]$ , and the residue field of  $K((t))$  is  $K$ , since this is the residue field of  $K(t)$  with the absolute value  $||_t$ .

A non-archimedean absolute value  $||$  on a field  $F$  will be said to be *discrete* if there exists some  $\delta \in (0, 1)$  such that  $a \in F$  and  $|a| \neq 1$  implies either  $|a| < 1 - \delta$  or  $|a| > 1 + \delta$ . (This situation cannot arise for archimedean absolute values.)

A non-archimedean absolute value need not be discrete, but the examples of non-archimedean absolute values which we have given are all discrete.

**Lemma 12** *Let  $F$  be a field with a nontrivial non-archimedean absolute value  $||$ , and let  $R$  and  $M$  be the corresponding valuation ring and valuation ideal. Then the absolute value is discrete if and only if  $M$  is a principal ideal. In this case the only nontrivial proper ideals of  $R$  are the powers  $M^k$  ( $k = 1, 2, \dots$ ).*

*Proof* Suppose first that the absolute value  $||$  is discrete and put  $\mu = \sup_{a \in M} |a|$ . Then  $0 < \mu < 1$  and the supremum is attained, since  $|a_n| \rightarrow \mu$  implies  $|a_{n+1}a_n^{-1}| \rightarrow 1$ . Thus  $\mu = |\pi|$  for some  $\pi \in M$ . For any  $a \in M$  we have  $|a\pi^{-1}| \leq 1$  and hence  $a = \pi a'$ , where  $a' \in R$ . Thus  $M$  is a principal ideal with generating element  $\pi$ .

Suppose next that  $M$  is a principal ideal with generating element  $\pi$ . If  $|a| < 1$ , then  $a \in M$ . Thus  $a = \pi a'$ , where  $a' \in R$ , and hence  $|a| \leq |\pi|$ . Similarly if  $|a| > 1$ , then  $a^{-1} \in M$ . Thus  $|a^{-1}| \leq |\pi|$  and hence  $|a| \geq |\pi|^{-1}$ . This proves that the absolute value is discrete.

We now show that, for any nonzero  $a \in M$ , there is a positive integer  $k$  such that  $|a| = |\pi|^k$ . In fact we can choose  $k$  so that

$$|\pi|^{k+1} < |a| \leq |\pi|^k.$$

Then  $|\pi| < |a\pi^{-k}| \leq 1$ , which implies  $|a\pi^{-k}| = 1$  and hence  $|a| = |\pi|^k$ . Thus the value group of the valued field  $F$  is the infinite cyclic group generated by  $|\pi|$ . The final statement of the lemma follows immediately.  $\square$

It is clear that if an absolute value  $|\cdot|$  on a field  $F$  is discrete, then its extension to the completion  $\bar{F}$  of  $F$  is also discrete. Moreover, if  $\pi$  is a generating element for the valuation ideal of  $F$ , then it is also a generating element for the valuation ideal of  $\bar{F}$ .

Suppose now that not only is  $M = (\pi)$  a principal ideal, but the residue field  $k = R/M$  is finite. Then there exists a finite set  $S \subseteq R$ , with the same cardinality as  $k$ , such that for each  $a \in R$  there is a unique  $\alpha \in S$  for which  $|\alpha - a| < 1$ . Since the elements of  $k$  are the cosets  $\alpha + M$ , where  $\alpha \in S$ , we call  $S$  a *set of representatives* in  $R$  of the residue field. It is convenient to choose  $\alpha = 0$  as the representative for  $M$  itself.

Under these hypotheses a rather explicit representation for the elements of the valued field can be derived:

**Proposition 13** *Let  $F$  be a field with a non-archimedean absolute value  $|\cdot|$ , and let  $R$  and  $M$  be the corresponding valuation ring and valuation ideal. Suppose the absolute value is discrete, i.e.  $M = (\pi)$  is a principal ideal. Suppose also that the residue field  $k = R/M$  is finite, and let  $S \subseteq R$  be a set of representatives of  $k$  with  $0 \in S$ .*

*Then for each  $a \in F$  there exists a unique bi-infinite sequence  $(\alpha_n)_{n \in \mathbb{Z}}$ , where  $\alpha_n \in S$  for all  $n \in \mathbb{Z}$  and  $\alpha_n \neq 0$  for at most finitely many  $n < 0$ , such that*

$$a = \sum_{n \in \mathbb{Z}} \alpha_n \pi^n.$$

*If  $N$  is the least integer  $n$  such that  $\alpha_n \neq 0$ , then  $|a| = |\pi|^N$ . In particular,  $a \in R$  if and only if  $\alpha_n = 0$  for all  $n < 0$ .*

*If  $F$  is complete then, for any such bi-infinite sequence  $(\alpha_n)$ , the series  $\sum_{n \in \mathbb{Z}} \alpha_n \pi^n$  is convergent with sum  $a \in F$ .*

*Proof* Suppose  $a \in F$  and  $a \neq 0$ . Then  $|a| = |\pi|^N$  for some  $N \in \mathbb{Z}$  and hence  $|a\pi^{-N}| = 1$ . There is a unique  $\alpha_N \in S$  such that  $|a\pi^{-N} - \alpha_N| < 1$ . Then  $|\alpha_N| = 1$ ,  $|a\pi^{-N} - \alpha_N| \leq |\pi|$  and

$$a\pi^{-N} = \alpha_N + a_1\pi,$$

where  $a_1 \in R$ . Similarly there is a unique  $\alpha_{N+1} \in S$  such that

$$a_1 = \alpha_{N+1} + a_2\pi,$$

where  $a_2 \in R$ . Continuing in this way we obtain, for any positive integer  $n$ ,

$$a = \alpha_N \pi^N + \alpha_{N+1} \pi^{N+1} + \cdots + \alpha_{N+n} \pi^{N+n} + a_{n+1} \pi^{N+n+1},$$

where  $\alpha_N, \alpha_{N+1}, \dots, \alpha_{N+n} \in S$  and  $a_{n+1} \in R$ . Since  $|a_{n+1}\pi^{N+n+1}| \rightarrow 0$  as  $n \rightarrow \infty$ , the series  $\sum_{n \geq N} \alpha_n \pi^n$  converges with sum  $a$ .

On the other hand, it is clear that if  $a = \sum_{n \geq N} \alpha_n \pi^n$ , where  $\alpha_n \in S$  and  $\alpha_N \neq 0$ , then the coefficients  $\alpha_n$  must be determined in the above way.

If  $F$  is complete then, by Corollary 11, any series  $\sum_{n \geq N} \alpha_n \pi^n$  is convergent, since  $|\alpha_n \pi^n| \rightarrow 0$  as  $n \rightarrow \infty$ .  $\square$

**Corollary 14** *Every  $a \in \mathbb{Q}_p$  can be uniquely expressed in the form*

$$a = \sum_{n \in \mathbb{Z}} \alpha_n p^n,$$

where  $\alpha_n \in \{0, 1, \dots, p-1\}$  and  $\alpha_n \neq 0$  for at most finitely many  $n < 0$ . Conversely, any such series is convergent with sum  $a \in \mathbb{Q}_p$ . Furthermore  $a \in \mathbb{Z}_p$  if and only if  $\alpha_n = 0$  for all  $n < 0$ .

Thus we have now arrived at Hensel's starting-point. It is not difficult to show that if  $a = \sum_{n \in \mathbb{Z}} \alpha_n p^n \in \mathbb{Q}_p$ , then actually  $a \in \mathbb{Q}$  if and only if the sequence of coefficients  $(\alpha_n)$  is *eventually periodic*, i.e. there exist integers  $h > 0$  and  $m$  such that  $\alpha_{n+h} = \alpha_n$  for all  $n \geq m$ .

From Corollary 14 we can deduce again that the ring  $\mathbb{Z}$  of ordinary integers is dense in the ring  $\mathbb{Z}_p$  of  $p$ -adic integers. For, if

$$a = \sum_{n \geq 0} \alpha_n p^n \in \mathbb{Z}_p,$$

where  $\alpha_n \in \{0, 1, \dots, p-1\}$ , then

$$a_k = \sum_{n=0}^k \alpha_n p^n \in \mathbb{Z}$$

and  $|a - a_k| < p^{-k}$ .

## 5 Hensel's Lemma

The analogy between  $p$ -adic absolute values and ordinary absolute values suggests that methods well-known in analysis may be applied also to arithmetic problems. We will illustrate this by showing how Newton's method for finding the real or complex roots of an equation can also be used to find  $p$ -adic roots. In fact the ultrametric inequality makes it possible to establish a stronger convergence criterion than in the classical case. The following proposition is modestly known as 'Hensel's lemma'.

**Proposition 15** *Let  $F$  be a field with a complete non-archimedean absolute value  $|\cdot|$  and let  $R$  be its valuation ring. Let*

$$f(x) = c_n x^n + c_{n-1} x^{n-1} + \dots + c_0$$

be a polynomial with coefficients  $c_0, \dots, c_n \in R$  and let

$$f_1(x) = nc_n x^{n-1} + (n-1)c_{n-1}x^{n-2} + \dots + c_1$$

be its formal derivative. If  $|f(a_0)| < |f_1(a_0)|^2$  for some  $a_0 \in R$ , then the equation  $f(a) = 0$  has a unique solution  $a \in R$  such that  $|a - a_0| < |f_1(a_0)|$ .

*Proof* We consider first the existence of  $a$  and postpone discussion of its uniqueness. Put

$$\sigma := |f_1(a_0)| > 0, \quad \theta_0 := \sigma^{-2}|f(a_0)| < 1,$$

and let  $D_\theta$  denote the set

$$\{a \in R : |f_1(a)| = \sigma, |f(a)| \leq \theta \sigma^2\}.$$

Thus  $a_0 \in D_{\theta_0}$  and  $D_{\theta'} \subseteq D_\theta$  if  $\theta' \leq \theta$ . We are going to show that, if  $\theta \in (0, 1)$ , then the 'Newton' map

$$Ta = a^* := a - f(a)/f_1(a)$$

maps  $D_\theta$  into  $D_{\theta^2}$ .

We can write

$$f(x+y) = f(x) + f_1(x)y + \dots + f_n(x)y^n,$$

where  $f_1(x)$  has already been defined and  $f_2(x), \dots, f_n(x)$  are also polynomials with coefficients from  $R$ . We substitute

$$x = a, y = b := -f(a)/f_1(a),$$

where  $a \in D_\theta$ . Then  $|f_j(a)| \leq 1$ , since  $a \in R$  and  $f_j(x) \in R[x]$  ( $j = 1, \dots, n$ ). Furthermore

$$|b| = \sigma^{-1}|f(a)| \leq \theta \sigma < \sigma.$$

Thus  $b \in R$ . Since  $f(a) + f_1(a)b = 0$ , it follows that  $a^* = a + b$  satisfies

$$|f(a^*)| \leq \max_{2 \leq j \leq n} |f_j(a)b^j| \leq |b|^2 = \sigma^{-2}|f(a)|^2 \leq \theta^2 \sigma^2.$$

Similarly, since  $f_1(a+b) - f_1(a)$  can be written as a polynomial in  $b$  with coefficients from  $R$  and with no constant term,

$$|f_1(a+b) - f_1(a)| \leq |b| < \sigma = |f_1(a)|$$

and hence  $|f_1(a^*)| = \sigma$ . This completes the proof that  $TD_\theta \subseteq D_{\theta^2}$ .

Now put  $a_k = T^k a_0$ , so that

$$a_{k+1} - a_k = -f(a_k)/f_1(a_k).$$

It follows by induction from what we have proved that

$$|f'(a_k)| \leq \theta_0^{2^k} \sigma^2.$$

Since  $\theta_0 < 1$  and  $|a_{k+1} - a_k| = \sigma^{-1} |f'(a_k)|$ , this shows that  $\{a_k\}$  is a fundamental sequence. Hence, since  $F$  is complete,  $a_k \rightarrow a$  for some  $a \in R$ . Evidently  $f(a) = 0$  and  $|f_1(a)| = \sigma$ . Since, for every  $k \geq 1$ ,

$$|a_k - a_0| \leq \max_{1 \leq j \leq k} |a_j - a_{j-1}| \leq \theta_0 \sigma,$$

we also have  $|a - a_0| \leq \theta_0 \sigma < \sigma$ .

To prove uniqueness, assume  $f(\tilde{a}) = 0$  for some  $\tilde{a} \neq a$  such that  $|\tilde{a} - a_0| < \sigma$ . If we put  $b = \tilde{a} - a$ , then

$$0 = f(\tilde{a}) - f(a) = f_1(a)b + \cdots + f_n(a)b^n.$$

From  $b = \tilde{a} - a_0 - (a - a_0)$  we obtain  $|b| < \sigma$ . Since  $b \neq 0$  and  $|f_j(a)| \leq 1$ , it follows that, for  $j \geq 2$ ,

$$|f_j(a)b^j| \leq |b|^2 < \sigma |b| = |f_1(a)b|.$$

But this implies

$$|f(\tilde{a}) - f(a)| = |f_1(a)b| > 0,$$

which is a contradiction.  $\square$

As an application of Proposition 15 we will determine which elements of the field  $\mathbb{Q}_p$  of  $p$ -adic numbers are squares. Since  $b = a^2$  implies  $b = p^{2v}b'$ , where  $v \in \mathbb{Z}$  and  $|b'|_p = 1$ , we may restrict attention to the case  $|b|_p = 1$ .

**Proposition 16** Suppose  $b \in \mathbb{Q}_p$  and  $|b|_p = 1$ .

If  $p \neq 2$ , then  $b = a^2$  for some  $a \in \mathbb{Q}_p$  if and only if  $|b - a_0^2|_p < 1$  for some  $a_0 \in \mathbb{Z}$ .

If  $p = 2$ , then  $b = a^2$  for some  $a \in \mathbb{Q}_2$  if and only if  $|b - 1|_2 \leq 2^{-3}$ .

*Proof* Suppose first that  $p \neq 2$ . If  $b = a^2$  for some  $a \in \mathbb{Q}_p$ , then  $|a|_p = 1$  and  $|a - a_0|_p < 1$  for some  $a_0 \in \mathbb{Z}$ , since  $\mathbb{Z}$  is dense in  $\mathbb{Z}_p$ . Hence  $|a_0|_p = 1$  and

$$|b - a_0^2|_p = |a - a_0|_p |a + a_0|_p \leq |a - a_0|_p < 1.$$

Conversely, suppose  $|b - a_0^2|_p < 1$  for some  $a_0 \in \mathbb{Z}$ . Then  $|a_0^2|_p = 1$  and so  $|a_0|_p = 1$ . In Proposition 15 take  $F = \mathbb{Q}_p$  and  $f(x) = x^2 - b$ . The hypotheses of the proposition are satisfied, since  $|f(a_0)|_p < 1$  and  $|f_1(a_0)|_p = |2a_0|_p = 1$ , and hence  $b = a^2$  for some  $a \in \mathbb{Q}_p$ .

Suppose next that  $p = 2$ . If  $b = a^2$  for some  $a \in \mathbb{Q}_2$ , then  $|a|_2 = 1$  and  $|a - a_0|_2 \leq 2^{-3}$  for some  $a_0 \in \mathbb{Z}$ , since  $\mathbb{Z}$  is dense in  $\mathbb{Z}_2$ . Hence  $|a_0|_2 = 1$  and

$$|b - a_0^2|_2 = |a - a_0|_2 |a + a_0|_2 \leq |a - a_0|_2 \leq 2^{-3}.$$

Since  $a_0$  is odd, we have  $a_0 \equiv \pm 1 \pmod{4}$  and  $a_0^2 \equiv 1 \pmod{8}$ . Hence

$$|b - 1|_2 \leq \max\{|b - a_0^2|_2, |a_0^2 - 1|_2\} \leq 2^{-3}.$$

Conversely, suppose  $|b - 1|_2 \leq 2^{-3}$ . In Proposition 15 take  $F = \mathbb{Q}_2$  and  $f(x) = x^2 - b$ . The hypotheses of the proposition are satisfied, since  $|f(1)|_2 \leq 2^{-3}$  and  $|f_1(1)|_2 = 2^{-1}$ , and hence  $b = a^2$  for some  $a \in \mathbb{Q}_2$ .  $\square$

**Corollary 17** *Let  $b$  be an integer not divisible by the prime  $p$ .*

*If  $p \neq 2$ , then  $b = a^2$  for some  $a \in \mathbb{Q}_p$  if and only if  $b$  is a quadratic residue mod  $p$ .*

*If  $p = 2$ , then  $b = a^2$  for some  $a \in \mathbb{Q}_2$  if and only if  $b \equiv 1 \pmod{8}$ .*

It follows from Corollary 17 that  $\mathbb{Q}_p$  cannot be given the structure of an ordered field. For, if  $p$  is odd, then  $1 - p = a^2$  for some  $a \in \mathbb{Q}_p$  and hence

$$a^2 + 1 + \cdots + 1 = 0,$$

where there are  $p - 1$  1's. Similarly, if  $p = 2$ , then  $1 - 2^3 = a^2$  for some  $a \in \mathbb{Q}_2$  and the same relation holds with 7 1's.

Suppose again that  $F$  is a field with a complete non-archimedean absolute value  $|\cdot|$ . Let  $R$  and  $M$  be the corresponding valuation ring and valuation ideal, and let  $k = R/M$  be the residue field. For any  $a \in R$  we will denote by  $\bar{a}$  the corresponding element  $a + M$  of  $k$ , and for any polynomial

$$f(x) = c_n x^n + c_{n-1} x^{n-1} + \cdots + c_0$$

with coefficients  $c_0, \dots, c_n \in R$ , we will denote by

$$\bar{f}(x) = \bar{c}_n x^n + \bar{c}_{n-1} x^{n-1} + \cdots + \bar{c}_0$$

the polynomial whose coefficients are the corresponding elements of  $k$ .

The hypotheses of Proposition 15 are certainly satisfied if  $|f(a_0)| < 1 = |f_1(a_0)|$ . In this case Proposition 15 says that if

$$\bar{f}(x) = (x - \bar{a}_0)\bar{h}_0(x),$$

where  $a_0 \in R$ ,  $h_0(x) \in R[x]$  and  $h_0(a_0) \notin M$ , then

$$f(x) = (x - a)h(x),$$

where  $a - a_0 \in M$ , and  $h(x) \in R[x]$ . In other words, the factorization of  $\bar{f}(x)$  in  $k[x]$  can be 'lifted' to a factorization of  $f(x)$  in  $R[x]$ . This form of Hensel's lemma can be generalized to factorizations where neither factor is linear, and the result is again known as Hensel's lemma!

**Proposition 18** *Let  $F$  be a field with a complete non-archimedean absolute value  $|\cdot|$ . Let  $R$  and  $M$  be the valuation ring and valuation ideal of  $F$ , and  $k = R/M$  the residue field.*

*Let  $f \in R[x]$  be a polynomial with coefficients in  $R$  and suppose there exist relatively prime polynomials  $\phi, \psi \in k[x]$ , with  $\phi$  monic and  $\partial(\phi) > 0$ , such that  $\bar{f} = \phi\psi$ .*

*Then there exist polynomials  $g, h \in R[x]$ , with  $g$  monic and  $\partial(g) = \partial(\phi)$ , such that  $\bar{g} = \phi$ ,  $\bar{h} = \psi$  and  $f = gh$ .*

*Proof* Put  $n = \partial(f)$  and  $m = \partial(\phi)$ . Then  $\partial(\psi) = \partial(\bar{f}) - \partial(\phi) \leq n - m$ . There exist polynomials  $g_1, h_1 \in R[x]$ , with  $g_1$  monic,  $\partial(g_1) = m$  and  $\partial(h_1) \leq n - m$ , such that  $\bar{g}_1 = \phi, \bar{h}_1 = \psi$ . Since  $\phi, \psi$  are relatively prime, there exist polynomials  $\chi, \omega \in k[x]$  such that

$$\chi\phi + \omega\psi = 1,$$

and there exist polynomials  $u, v \in R[x]$  such that  $\bar{u} = \chi, \bar{v} = \omega$ . Thus

$$f - g_1 h_1 \in M[x], \quad u g_1 + v h_1 - 1 \in M[x].$$

If  $f = g_1 h_1$ , there is nothing more to do. Otherwise, let  $\pi$  be the coefficient of  $f - g_1 h_1$  or of  $u g_1 + v h_1 - 1$  which has maximum absolute value. Then

$$f - g_1 h_1 \in \pi R[x], \quad u g_1 + v h_1 - 1 \in \pi R[x].$$

We are going to construct inductively polynomials  $g_j, h_j \in R[x]$  such that

- (i)  $\bar{g}_j = \phi, \bar{h}_j = \psi$ ;
- (ii)  $g_j$  is monic and  $\partial(g_j) = m, \partial(h_j) \leq n - m$ ;
- (iii)  $g_j - g_{j-1} \in \pi^{j-1} R[x], h_j - h_{j-1} \in \pi^{j-1} R[x]$ ;
- (iv)  $f - g_j h_j \in \pi^j R[x]$ .

This holds already for  $j = 1$  with  $g_0 = h_0 = 0$ . Assume that, for some  $k \geq 2$ , it holds for all  $j < k$  and put  $f - g_j h_j = \pi^j \ell_j$ , where  $\ell_j \in R[x]$ . Since  $g_1$  is monic, the Euclidean algorithm provides polynomials  $q_k, r_k \in R[x]$  such that

$$\ell_{k-1} v = q_k g_1 + r_k, \quad \partial(r_k) < \partial(g_1) = m.$$

Let  $w_k \in R[x]$  be a polynomial of minimal degree such that all coefficients of  $\ell_{k-1} u + q_k h_1 - w_k$  have absolute value at most  $|\pi|$ . Then

$$w_k g_1 + r_k h_1 - \ell_{k-1} = (u g_1 + v h_1 - 1) \ell_{k-1} - (\ell_{k-1} u + q_k h_1 - w_k) g_1 \in \pi R[x].$$

We will show that  $\partial(w_k) \leq n - m$ . Indeed otherwise

$$\partial(w_k g_1) > n \geq \partial(r_k h_1 - \ell_{k-1})$$

and hence, since  $g_1$  is monic,  $w_k g_1 + r_k h_1 - \ell_{k-1}$  has the same leading coefficient as  $w_k$ . Consequently the leading coefficient of  $w_k$  is in  $\pi R$ . Thus the polynomial obtained from  $w_k$  by omitting the term of highest degree satisfies the same requirements as  $w_k$ , which is a contradiction.

If we put

$$g_k = g_{k-1} + \pi^{k-1} r_k, \quad h_k = h_{k-1} + \pi^{k-1} w_k,$$

then (i)–(iii) are evidently satisfied for  $j = k$ . Moreover

$$f - g_k h_k = -\pi^{k-1} (w_k g_{k-1} + r_k h_{k-1} - \ell_{k-1}) - \pi^{2k-2} r_k w_k$$

and

$$\begin{aligned} & w_k g_{k-1} + r_k h_{k-1} - \ell_{k-1} \\ &= w_k g_1 + r_k h_1 - \ell_{k-1} + w_k (g_{k-1} - g_1) + r_k (h_{k-1} - h_1) \in \pi R[x]. \end{aligned}$$

Hence also (iv) is satisfied for  $j = k$ .

Put

$$g_j(x) = x^m + \sum_{i=0}^{m-1} \alpha_i^{(j)} x^i, \quad h_j(x) = \sum_{i=0}^{n-m} \beta_i^{(j)} x^i.$$

By (iii), the sequences  $(\alpha_i^{(j)})$  and  $(\beta_i^{(j)})$  are fundamental sequences for each  $i$  and hence, since  $F$  is complete, there exist  $\alpha_i, \beta_i \in R$  such that

$$\alpha_i^{(j)} \rightarrow \alpha_i, \beta_i^{(j)} \rightarrow \beta_i \text{ as } j \rightarrow \infty.$$

If

$$g(x) = x^m + \sum_{i=0}^{m-1} \alpha_i x^i, \quad h(x) = \sum_{i=0}^{n-m} \beta_i x^i,$$

then, for each  $j \geq 1$ ,

$$g - g_j \in \pi^j R[x], \quad h - h_j \in \pi^j R[x].$$

Since

$$f - gh = f - g_j h_j - (g - g_j)h - g_j(h - h_j),$$

it follows that  $f - gh \in \pi^j R[x]$  for each  $j \geq 1$ . Hence  $f = gh$ . It is obvious that  $g$  and  $h$  have the other required properties.  $\square$

As an application of this form of Hensel's lemma we prove

**Proposition 19** *Let  $F$  be a field with a complete non-archimedean absolute value  $|\cdot|$  and let*

$$f(t) = c_n t^n + c_{n-1} t^{n-1} + \cdots + c_0 \in F[t].$$

*If  $c_0 c_n \neq 0$  and, for some  $m$  such that  $0 < m < n$ ,*

$$|c_0| \leq |c_m|, \quad |c_n| \leq |c_m|,$$

*with at least one of the two inequalities strict, then  $f$  is reducible over  $F$ .*

*Proof* Suppose first that  $|c_0| < |c_m|$  and  $|c_n| \leq |c_m|$ . Evidently we may choose  $m$  so that  $|c_m| = \max_{0 \leq i < n} |c_i|$  and  $|c_i| < |c_m|$  for  $0 \leq i < m$ . By multiplying  $f$  by  $c_m^{-1}$  we may further assume that, if  $R$  is the valuation ring of  $F$ , then  $f(t) \in R[t]$ ,  $c_m = 1$  and  $|c_i| < 1$  for  $0 \leq i < m$ . Hence

$$\bar{f}(t) = t^m (\bar{c}_n t^{n-m} + \bar{c}_{n-1} t^{n-m-1} + \cdots + 1).$$

Since the two factors are relatively prime, it follows from Proposition 18 that  $f$  is reducible.

If  $|c_n| < |c_m|$  and  $|c_0| \leq |c_m|$ , then the same argument also applies to the polynomial  $t^n f(t^{-1})$ .  $\square$

Proposition 19 shows that if a quadratic polynomial  $at^2 + bt + c$  is irreducible, then  $|b| \leq \max\{|a|, |c|\}$ , with strict inequality if  $|a| \neq |c|$ . Proposition 19 will now be used to extend an absolute value on a given field to a finite extension of that field.

**Proposition 20** *Let  $F$  be a field with a complete non-archimedean absolute value  $|\cdot|$ . If the field  $E$  is a finite extension of  $F$ , then the absolute value on  $F$  can be extended to an absolute value on  $E$ .*

*Proof* We will not only show that an extension of the absolute value exists, but we will provide an explicit expression for it.

Regard  $E$  as a vector space over  $F$  of finite dimension  $n$ , and with any  $a \in E$  associate the linear transformation  $L_a: E \rightarrow E$  defined by  $L_a(x) = ax$ . Then  $\det L_a \in F$  and we claim that an extended absolute value is given by the formula

$$|a| = |\det L_a|^{1/n}.$$

Evidently  $|a| \geq 0$ , and equality holds only if  $a = 0$ , since  $ax = 0$  for some  $x \neq 0$  implies  $a = 0$ . Furthermore  $|ab| = |a||b|$ , since  $L_{ab} = L_a L_b$  and hence  $\det L_{ab} = (\det L_a)(\det L_b)$ . If  $a \in F$ , then  $L_a = aI_n$  and hence the proposed absolute value coincides with the original absolute value on  $F$ . It only remains to show that

$$|a - b| \leq \max(|a|, |b|) \quad \text{for all } a, b \in F.$$

In fact we may suppose  $|a| \leq |b|$  and then, by dividing by  $b$ , we see that it is sufficient to show that  $0 < |a| \leq 1$  implies  $|1 - a| \leq 1$ .

To simplify notation, write  $A = L_a$  and let

$$f(t) = \det(tI - A) = t^n + c_{n-1}t^{n-1} + \cdots + c_0$$

be the characteristic polynomial of  $A$ . Then  $c_i \in F$  for all  $i$  and  $c_0 = (-1)^n \det A$ . Let  $g(t)$  be the monic polynomial in  $F[t]$  of least positive degree such that  $g(a) = 0$ . Then  $g(t)$  is irreducible, since the field  $E$  has no zero divisors. Evidently  $g(t)$  is also the minimal polynomial of  $A$ . But, for an arbitrary linear transformation of an  $n$ -dimensional vector space, the characteristic polynomial divides the  $n$ -th power of the minimal polynomial (see M. Deuring, *Algebra*, p.4). It follows in the present case that  $f(t) = g(t)^r$  for some positive integer  $r$ .

Suppose

$$g(t) = t^m + b_{m-1}t^{m-1} + \cdots + b_0$$

and let  $a \in E$  satisfy  $|a| \leq 1$  with respect to the proposed absolute value. Then  $|c_0| = |\det A| \leq 1$  and hence, since  $b_0^r = c_0$ ,  $|b_0| \leq 1$ . Since  $g$  is irreducible, it follows from Proposition 19 that  $|b_j| \leq 1$  for all  $j$ . Since

$$g(1) = 1 + b_{m-1} + \cdots + b_0,$$

this implies  $|g(1)| \leq 1$  and hence  $|f(1)| \leq 1$ . Since  $f(1) = \det(I - A)$ , this proves that  $|1 - a| \leq 1$ .  $\square$

Finally we show that there is no other extension to  $E$  of the given absolute value on  $F$  besides the one constructed in the proof of Proposition 20.

**Proposition 21** *Let  $F$  be a complete field with respect to the absolute value  $|\cdot|$  and let the field  $E$  be a finite extension of  $F$ . Then there is at most one extension of the absolute value on  $F$  to an absolute value on  $E$ , and  $E$  is necessarily complete with respect to the extended absolute value.*

*Proof* Let  $e_1, \dots, e_n$  be a basis for  $E$ , regarded as a vector space over  $F$ . Then any  $a \in E$  can be uniquely expressed in the form

$$a = \alpha_1 e_1 + \dots + \alpha_n e_n,$$

where  $\alpha_1, \dots, \alpha_n \in F$ . By Lemma 7, for any extended absolute value there exist positive real numbers  $\sigma, \mu$  such that

$$\sigma|a| \leq \max_i |\alpha_i| \leq \mu|a| \quad \text{for every } a \in E.$$

It follows at once that  $E$  is complete. For if  $a^{(k)}$  is a fundamental sequence, then  $\alpha_i^{(k)}$  is a fundamental sequence in  $F$  for  $i = 1, \dots, n$ . Since  $F$  is complete, there exist  $\alpha_i \in F$  such that  $\alpha_i^{(k)} \rightarrow \alpha_i$  ( $i = 1, \dots, n$ ) and then  $a^{(k)} \rightarrow a$ , where  $a = \alpha_1 e_1 + \dots + \alpha_n e_n$ .

It will now be shown that there is at most one extension to  $E$  of the absolute value on  $F$ . Since we saw in §4 that the trivial absolute value on  $E$  is the only extension of the trivial absolute value on  $F$ , we may assume that the given absolute value on  $F$  is nontrivial. For a fixed  $a \in E$ , consider the powers  $a, a^2, \dots$ . For each  $k$  we can write

$$a^k = \alpha_1^{(k)} e_1 + \dots + \alpha_n^{(k)} e_n.$$

Since  $|a| < 1$  if and only if  $|a^k| \rightarrow 0$ , it follows from the remarks at the beginning of the proof that  $|a| < 1$  if and only if  $|\alpha_i^{(k)}| \rightarrow 0$  ( $i = 1, \dots, n$ ). This condition is independent of the absolute value on  $E$ . Thus if there exist two absolute values,  $|\cdot|_1$  and  $|\cdot|_2$ , which extend the absolute value on  $F$ , then  $|a|_1 < 1$  if and only if  $|a|_2 < 1$ . Hence, by Proposition 3, there exists a positive real number  $\rho$  such that

$$|a|_2 = |a|_1^\rho \quad \text{for every } a \in E.$$

In fact  $\rho = 1$ , since for some  $a \in F$  we have  $|a|_2 = |a|_1 > 1$ . □

## 6 Locally Compact Valued Fields

We prove first a theorem of Ostrowski (1918):

**Theorem 22** *A complete archimedean valued field  $F$  is (isomorphic to) either the real field  $\mathbb{R}$  or the complex field  $\mathbb{C}$ , and its absolute value is equivalent to the usual absolute value.*

*Proof* Since the valuation on it is archimedean, the field  $F$  has characteristic 0 and thus contains  $\mathbb{Q}$ . Since an archimedean absolute value on  $\mathbb{Q}$  is equivalent to the usual absolute value, by replacing the given absolute value on  $F$  by an equivalent one we may assume that it reduces to the usual absolute value on  $\mathbb{Q}$ . Since the valuation on  $F$

is complete, it now follows that  $F$  contains (a copy of)  $\mathbb{R}$  and that the absolute value on  $F$  reduces to the usual absolute value on  $\mathbb{R}$ . If  $F$  contains an element  $i$  such that  $i^2 = -1$ , then  $F$  contains (a copy of)  $\mathbb{C}$  and, by Proposition 21, the absolute value on  $F$  reduces to the usual absolute value on  $\mathbb{C}$ .

We now show that if  $a \in F$  and  $|a| < 1$ , then  $1 - a$  is a square in  $F$ . Let  $B$  be the set of all  $x \in F$  such that  $|x| \leq |a|$  and, for any  $x \in B$ , put

$$Tx = (x^2 + a)/2.$$

Then also  $Tx \in B$ , since

$$|Tx| \leq (|x|^2 + |a|)/2 \leq (|a|^2 + |a|)/2 \leq |a|.$$

Moreover, the map  $T$  is a contraction since, for all  $x, y \in B$ ,

$$|Tx - Ty| = |x^2 - y^2|/2 = |x - y||x + y|/2 \leq |a||x - y|.$$

Since  $F$  is complete and  $B$  is a closed subset of  $F$ , it follows from the contraction principle (Proposition I.26) that the map  $T$  has a fixed point  $\bar{x} \in B$ . Evidently  $\bar{x} = (\bar{x}^2 + a)/2$  and

$$1 - a = 1 - 2\bar{x} + \bar{x}^2 = (1 - \bar{x})^2.$$

We show next that, if the polynomial  $t^2 + 1$  does not have a root in  $F$ , then the valuation on  $F$  can be extended to the field  $E = F(i)$ , where  $i^2 = -1$ . Each  $\gamma \in E$  has a unique representation  $\gamma = a + ib$ , where  $a, b \in F$ . We claim that  $|\gamma| = \sqrt{|a|^2 + |b|^2}$  is an extension to  $E$  of the given valuation on  $F$ .

The only part of this claim which is not easily established is the triangle inequality. To prove it, we need only show that

$$|1 + \gamma| \leq 1 + |\gamma| \quad \text{for every } \gamma \in E.$$

That is, we need only show that

$$|(1 + a)^2 + b^2| \leq 1 + 2\sqrt{|a|^2 + |b|^2} + |a|^2 + |b|^2 \quad \text{for all } a, b \in F.$$

Since, by the triangle inequality in  $F$ ,

$$|(1 + a)^2 + b^2| \leq 1 + 2|a| + |a|^2 + |b|^2,$$

it is enough to show that

$$|a| \leq \sqrt{|a|^2 + |b|^2} \quad \text{for all } a, b \in F$$

or, since we may suppose  $a \neq 0$ ,

$$1 \leq |1 + c^2| \quad \text{for every } c \in F.$$

Assume, on the contrary, that  $|1 + c^2| < 1$  for some  $c \in F$ . Then, by the previous part of the proof,

$$-c^2 = 1 - (1 + c^2) = x^2 \quad \text{for some } x \in F.$$

Since  $c \neq 0$ , this implies that  $-1 = i^2$  for some  $i \in F$ , which is a contradiction.

Now  $E = F(i)$  contains  $\mathbb{C}$  and the absolute value on  $E$  reduces to the usual absolute value on  $\mathbb{C}$ . To prove the theorem it is enough to show that  $E = \mathbb{C}$ . For then  $\mathbb{R} \subseteq F \subseteq \mathbb{C}$  and  $F$  has dimension 1 or 2 as a vector space over  $\mathbb{R}$  according as  $i \notin F$  or  $i \in F$ .

Assume on the contrary that there exists  $\zeta \in E \setminus \mathbb{C}$ . Consider the function  $\varphi: \mathbb{C} \rightarrow \mathbb{R}$  defined by

$$\varphi(z) = |z - \zeta|$$

and put  $r = \inf_{z \in \mathbb{C}} \varphi(z)$ . Since  $\varphi(0) = |\zeta|$  and  $\varphi(z) > |\zeta|$  for  $|z| > 2|\zeta|$ , and since  $\varphi$  is continuous, the compact set  $\{z \in \mathbb{C}: |z| \leq 2|\zeta|\}$  contains a point  $w$  such that  $\varphi(w) = r$ .

Thus if we put  $\omega = \zeta - w$ , then  $\omega \neq 0$  and

$$0 < r = |\omega| \leq |\omega - z| \quad \text{for every } z \in \mathbb{C}.$$

We will show that  $|\omega - z| = r$  for every  $z \in \mathbb{C}$  such that  $|z| < r$ .

If  $\varepsilon = e^{2\pi i/n}$ , then

$$\omega^n - z^n = (\omega - z)(\omega - \varepsilon z) \cdots (\omega - \varepsilon^{n-1}z)$$

and hence

$$|\omega^n - z^n| \geq r^{n-1}|\omega - z|.$$

Thus  $|\omega - z| \leq r|1 - z^n/\omega^n|$ . Since  $|z| < |\omega|$ , by letting  $n \rightarrow \infty$  we obtain  $|\omega - z| \leq r$ . But this is possible only if  $|\omega - z| = r$ .

Thus if  $0 < |z| < r$ , then  $\omega$  may be replaced by  $\omega - z$ . It follows that  $|\omega - nz| = r$  for every positive integer  $n$ . Hence  $r \geq n|z| - r$ , which yields a contradiction for sufficiently large  $n$ .  $\square$

If a field  $F$  is locally compact with respect to an archimedean absolute value, then it is certainly complete and so, by Theorem 22, it is equivalent either to  $\mathbb{R}$  or to  $\mathbb{C}$  with the usual absolute value. It will now be shown that a field  $F$  is locally compact with respect to a non-archimedean absolute value if and only if it is a complete field of the type discussed in Proposition 13. It should be observed that a non-archimedean valued field  $F$  is locally compact if and only if its valuation ring  $R$  is compact, since then any closed ball in  $F$  is compact.

**Proposition 23** *Let  $F$  be a field with a non-archimedean absolute value  $|\cdot|$ . Then  $F$  is locally compact with respect to the topology induced by the absolute value if and only if the following three conditions are satisfied:*

- (i)  $F$  is complete,
- (ii) the absolute value  $|\cdot|$  is discrete,
- (iii) the residue field is finite.

*Proof* As we have just observed,  $F$  is locally compact if and only if its valuation ring  $R$  is compact. Moreover, since  $R$  is a subset of the metric space  $F$ , it is compact if and only if any sequence of elements of  $R$  has a convergent subsequence.

The field  $F$  is certainly complete if it is locally compact, since any fundamental sequence is bounded. If the residue field is infinite, then there exists an infinite sequence  $(a_k)$  of elements of  $R$  such that  $|a_k - a_j| = 1$  for  $j \neq k$ . Since the sequence  $(a_k)$  has no convergent subsequence,  $R$  is not compact. If the absolute value  $|\cdot|$  is not discrete, then there exists an infinite sequence  $(a_k)$  of elements of  $R$  with

$$|a_1| < |a_2| < \cdots$$

and  $|a_k| \rightarrow 1$  as  $k \rightarrow \infty$ . If  $k > j$ , then  $|a_k - a_j| = |a_k|$  and again the sequence  $(a_k)$  has no convergent subsequence. Thus the conditions (i)–(iii) are all necessary for  $F$  to be locally compact.

Suppose now that the conditions (i)–(iii) are all satisfied and let  $\sigma = (a_k)$  be a sequence of elements of  $R$ . In the notation of Proposition 13, let

$$a_k = \sum_{n \geq 0} \alpha_n^{(k)} \pi^n,$$

where  $\alpha_n^{(k)} \in S$ . Since  $S$  is finite, there exists  $\alpha_0 \in S$  such that  $\alpha_0^{(k)} = \alpha_0$  for infinitely many  $a_k \in \sigma$ . If  $\sigma_0$  is the subsequence of  $\sigma$  containing those  $a_k$  for which  $\alpha_0^{(k)} = \alpha_0$ , then there exists  $\alpha_1 \in S$  such that  $\alpha_1^{(k)} = \alpha_1$  for infinitely many  $a_k \in \sigma_0$ . Similarly, if  $\sigma_1$  is the subsequence of  $\sigma_0$  containing those  $a_k$  for which  $\alpha_1^{(k)} = \alpha_1$ , then there exists  $\alpha_2 \in S$  such that  $\alpha_2^{(k)} = \alpha_2$  for infinitely many  $a_k \in \sigma_1$ . And so on. If  $a^{(j)} \in \sigma_j$ , then

$$a^{(j)} = \alpha_0 + \alpha_1 \pi + \cdots + \alpha_j \pi^j + \sum_{n \geq 0} \alpha_n(j) \pi^{j+1+n}.$$

But  $a = \sum_{n \geq 0} \alpha_n \pi^n \in F$ , since  $F$  is complete, and  $|a^{(j)} - a| \leq |\pi|^{j+1}$ . Thus the subsequence  $(a^{(j)})$  of  $\sigma$  converges to  $a$ .  $\square$

**Corollary 24** *The field  $\mathbb{Q}_p$  of  $p$ -adic numbers is locally compact, and the ring  $\mathbb{Z}_p$  of  $p$ -adic integers is compact.*

**Corollary 25** *If  $K$  is a finite field, then the field  $K((t))$  of all formal Laurent series is locally compact, and the ring  $K[[t]]$  of all formal power series is compact.*

We now show that all locally compact valued fields  $F$  with a non-archimedean absolute value can in fact be explicitly determined. It is convenient to treat the cases where  $F$  has prime characteristic and zero characteristic separately, since the arguments in the two cases are quite different.

**Lemma 26** *Let  $F$  be a locally compact valued field with a nontrivial valuation. A normed vector space  $E$  over  $F$  is locally compact if and only if it is finite-dimensional.*

*Proof* Suppose first that  $E$  is finite-dimensional over  $F$ . If  $e_1, \dots, e_n$  is a basis for the vector space  $E$ , then any  $a \in E$  can be uniquely represented in the form

$$a = \alpha_1 e_1 + \cdots + \alpha_n e_n,$$

where  $\alpha_1, \dots, \alpha_n \in F$ , and

$$\|a\|_0 = \max_{1 \leq i \leq n} |\alpha_i|$$

is a norm on  $E$ . Since the field  $F$  is locally compact, it is also complete. Hence, by Lemma 7, there exist positive real constants  $\sigma, \mu$  such that

$$\sigma \|a\|_0 \leq \|a\| \leq \mu \|a\|_0 \quad \text{for every } a \in E.$$

Consequently, if  $\{a_k\}$  is a bounded sequence of elements of  $E$  then, for each  $j \in \{1, \dots, n\}$ , the corresponding coefficients  $\{\alpha_{kj}\}$  form a bounded sequence of elements of  $F$ . Hence, since  $F$  is locally compact, there exists a subsequence  $\{a_{k_v}\}$  such that each of the sequences  $\{\alpha_{k_v j}\}$  converges in  $F$ , with limit  $\beta_j$  say ( $j = 1, \dots, n$ ). It follows that the subsequence  $\{a_{k_v}\}$  converges in  $E$  with limit  $b = \beta_1 e_1 + \dots + \beta_n e_n$ . Thus  $E$  is locally compact.

Suppose next that  $E$  is infinite-dimensional over  $F$ . Since the valuation on  $F$  is nontrivial, there exists  $\alpha \in F$  such that  $r = |\alpha|$  satisfies  $0 < r < 1$ . Let  $V$  be any finite-dimensional subspace of  $E$ , let  $u' \in E \setminus V$  and let

$$d = \inf_{v \in V} \|u' - v\|.$$

Since  $V$  is locally compact,  $d > 0$  and  $d = \|u' - v'\|$  for some  $v' \in V$ . Choose  $k \in \mathbb{Z}$  so that  $r^{k+1} < d \leq r^k$  and put  $w' = \alpha^{-k}(u' - v')$ . For any  $v \in V$ ,

$$\|\alpha^k v + v' - u'\| \geq d$$

and hence

$$\|w' - v\| \geq dr^{-k} > r.$$

On the other hand,

$$\|w'\| = dr^{-k} \leq 1.$$

We now define a sequence  $\{w_m\}$  of elements of  $E$  in the following way. Taking  $V = \{O\}$  we obtain a vector  $w_1$  with  $r < \|w_1\| \leq 1$ . Suppose we have defined  $w_1, \dots, w_m \in E$  so that, for  $1 \leq j \leq m$ ,  $\|w_j\| \leq 1$  and  $\|w_j - v_j\| > r$  for all  $v_j$  in the vector subspace  $V_{j-1}$  of  $E$  spanned by  $w_1, \dots, w_{j-1}$ . Then, taking  $V = V_m$ , we obtain a vector  $w_{m+1}$  such that  $\|w_{m+1}\| \leq 1$  and  $\|w_{m+1} - v_{m+1}\| > r$  for all  $v_{m+1} \in V_m$ . Thus the process can be continued indefinitely. Since  $\|w_m\| \leq 1$  for all  $m$  and  $\|w_m - w_j\| > r$  for  $1 \leq j < m$ , the bounded sequence  $\{w_m\}$  has no convergent subsequence. Thus  $E$  is not locally compact.  $\square$

**Proposition 27** *A non-archimedean valued field  $E$  with zero characteristic is locally compact if and only if, for some prime  $p$ ,  $E$  is isomorphic to a finite extension of the field  $\mathbb{Q}_p$  of  $p$ -adic numbers.*

*Proof* If  $E$  is a finite extension of the  $p$ -adic field  $\mathbb{Q}_p$  then, since  $\mathbb{Q}_p$  is locally compact, so also is  $E$  by Lemma 26.

Suppose on the other hand that  $E$  is a locally compact valued field with zero characteristic. Then  $\mathbb{Q} \subseteq E$ . By Proposition 23, the residue field  $k = R/M$  is finite and

thus has prime characteristic  $p$ . It follows from Proposition 4 that the restriction to  $\mathbb{Q}$  of the absolute value on  $E$  is (equivalent to) the  $p$ -adic absolute value. Hence, since  $E$  is necessarily complete,  $\mathbb{Q}_p \subseteq E$ . If  $E$  were infinite-dimensional as a vector space over  $\mathbb{Q}_p$  then, by Lemma 26, it would not be locally compact. Hence  $E$  is a finite extension of  $\mathbb{Q}_p$ .  $\square$

We consider next locally compact valued fields of prime characteristic.

**Proposition 28** *A valued field  $F$  with prime characteristic  $p$  is locally compact if and only if  $F$  is isomorphic to the field  $K((t))$  of formal Laurent series over a finite field  $K$  of characteristic  $p$ , with the absolute value defined in example (iv) of §1. The finite field  $K$  is the residue field of  $F$ .*

*Proof* We need only prove the necessity of the condition, since (Corollary 25) we have already established its sufficiency. Since  $F$  has prime characteristic, the absolute value on  $F$  is non-archimedean. Hence, by Proposition 23 and Lemma 12, the absolute value on  $F$  is discrete and the valuation ideal  $M$  is a principal ideal. Let  $\pi$  be a generating element for  $M$ . By Proposition 23 also, the residue field  $k = R/M$  is finite. Evidently the characteristic of  $k$  must also be  $p$ . Let  $q = p^f$  be the number of elements in  $k$ . Since  $F$  has characteristic  $p$ , for any  $a, b \in F$ ,

$$(b - a)^p = b^p - a^p$$

and hence, by induction,

$$(b - a)^{p^n} = b^{p^n} - a^{p^n} \quad \text{for all } n \geq 1.$$

The multiplicative group of  $k$  is a cyclic group of order  $q - 1$ . Choose  $a \in R$  so that  $a + M$  generates this cyclic group. Then  $|a^q - a| < 1$ . By what we have just proved,

$$a^{q^{n+1}} - a^{q^n} = (a^q - a)^{q^n},$$

and hence  $(a^{q^n})$  is a fundamental sequence, by Lemma 10. Since  $F$  is complete, by Proposition 23, it follows that  $a^{q^n} \rightarrow \alpha \in R$ . Moreover  $\alpha^q = \alpha$ , since

$$\lim_{n \rightarrow \infty} (a^{q^n})^q = \lim_{n \rightarrow \infty} a^{q^{n+1}},$$

and  $\alpha - a \in M$ , since  $a^{q^{n+1}} - a^{q^n} \in M$  for every  $n \geq 0$ . Hence  $\alpha \neq 0$  and  $\alpha^{q-1} = 1$ . Moreover  $\alpha^j \neq 1$  for  $1 \leq j < q - 1$ , since  $\alpha^j \equiv a^j \pmod{M}$ . It follows that the set  $S$  consisting of 0 and the powers  $1, \alpha, \dots, \alpha^{q-1}$  is a set of representatives in  $R$  of the residue field  $k$ .

Since  $F$  has characteristic  $p$ ,  $\alpha$  generates a finite subring  $K$  of  $R$ . In fact  $K$  is a field, since  $\beta^q = \beta$  for every  $\beta \in K$  and so  $\beta\beta^{q-2} = 1$  if  $\beta \neq 0$ . Since  $S \subseteq K$  and the polynomial  $x^q - x$  has at most  $q$  roots in  $K$ , we conclude that  $S = K$ . Thus  $K$  has  $q$  elements and is isomorphic to the residue field  $k$ .

Every element  $a$  of  $F$  has a unique representation

$$a = \sum_{n \in \mathbb{Z}} \alpha_n \pi^n,$$

where  $\pi$  is a generating element for the principal ideal  $M$ ,  $\alpha_n \in S$  and  $\alpha_n \neq 0$  for at most finitely many  $n < 0$ . The map

$$a' = \sum_{n \in \mathbb{Z}} \alpha_n t^n \rightarrow a = \sum_{n \in \mathbb{Z}} \alpha_n \pi^n$$

is a bijection of the field  $K((t))$  onto  $F$ . Since  $S$  is closed under addition this map preserves sums, and since  $S$  is also closed under multiplication it also preserves products. Finally, if  $N$  is the least integer such that  $\alpha_N \neq 0$ , then  $|a| = |\pi|^N$  and  $|a'| = \rho^{-N}$  for some fixed  $\rho > 1$ . Hence the map is an isomorphism of the valued field  $K((t))$  onto  $F$ .  $\square$

## 7 Further Remarks

Valued fields are discussed in more detail in the books of Cassels [1], Endler [3] and Ribenboim [5].

For still more forms of Hensel's lemma, see Ribenboim [6]. There are also generalizations to polynomials in several variables and to power series. The algorithmic implementation of Hensel's lemma is studied in von zur Gathen [4]. Newton's method for finding real or complex zeros is discussed in Stoer and Bulirsch [7], for example.

Proposition 20 continues to hold if the word 'complete' is omitted from its statement. However, the formula given in the proof of Proposition 20 defines an absolute value on  $E$  if and only if there is a *unique* extension of the absolute value on  $F$  to an absolute value on  $E$ ; see Viswanathan [8].

Ostrowski's Theorem 22 has been generalized by weakening the requirement  $|ab| = |a||b|$  to  $|ab| \leq |a||b|$ . Mazur (1938) proved that the only normed associative division algebras over  $\mathbb{R}$  are  $\mathbb{R}$ ,  $\mathbb{C}$  and  $\mathbb{H}$ , and that the only normed associative division algebra over  $\mathbb{C}$  is  $\mathbb{C}$  itself. An elegant functional-analytic proof of the latter result was given by Gelfand (1941). See Chapter 8 (by Koecher and Remmert) of Ebbinghaus *et al.* [2].

## 8 Selected References

- [1] J.W.S. Cassels, *Local fields*, Cambridge University Press, 1986.
- [2] H.-D. Ebbinghaus *et al.*, *Numbers*, English transl. of 2nd German ed. by H.L.S. Orde, Springer-Verlag, New York, 1990.
- [3] O. Endler, *Valuation theory*, Springer-Verlag, Berlin, 1972.
- [4] J. von zur Gathen, Hensel and Newton methods in valuation rings, *Math. Comp.* **42** (1984), 637–661.
- [5] P. Ribenboim, *The theory of classical valuations*, Springer-Verlag, New York, 1999.
- [6] P. Ribenboim, Equivalent forms of Hensel's lemma, *Exposition. Math.* **3** (1985), 3–24.
- [7] J. Stoer and R. Bulirsch, *Introduction to numerical analysis*, 3rd ed. (English transl.), Springer-Verlag, New York, 2002.
- [8] T.M. Viswanathan, A characterisation of Henselian valuations via the norm, *Bol. Soc. Brasil. Mat.* **4** (1973), 51–53.

## VII

# The Arithmetic of Quadratic Forms

We have already determined the integers which can be represented as a sum of two squares. Similarly, one may ask which integers can be represented in the form  $x^2 + 2y^2$  or, more generally, in the form  $ax^2 + 2bxy + cy^2$ , where  $a, b, c$  are given integers. The arithmetic theory of binary quadratic forms, which had its origins in the work of Fermat, was extensively developed during the 18th century by Euler, Lagrange, Legendre and Gauss. The extension to quadratic forms in more than two variables, which was begun by them and is exemplified by Lagrange's theorem that every positive integer is a sum of four squares, was continued during the 19th century by Dirichlet, Hermite, H.J.S. Smith, Minkowski and others. In the 20th century Hasse and Siegel made notable contributions. With Hasse's work especially it became apparent that the theory is more perspicuous if one allows the variables to be rational numbers, rather than integers. This opened the way to the study of quadratic forms over arbitrary fields, with pioneering contributions by Witt (1937) and Pfister (1965–67).

From this vast theory we focus attention on one central result, the *Hasse–Minkowski theorem*. However, we first study quadratic forms over an arbitrary field in the geometric formulation of Witt. Then, following an interesting approach due to Fröhlich (1967), we study quadratic forms over a *Hilbert field*.

## 1 Quadratic Spaces

The theory of quadratic spaces is simply another name for the theory of quadratic forms. The advantage of the change in terminology lies in its appeal to geometric intuition. It has in fact led to new results even at quite an elementary level. The new approach had its debut in a paper by Witt (1937) on the arithmetic theory of quadratic forms, but it is appropriate also if one is interested in quadratic forms over the real field or any other field.

For the remainder of this chapter we will restrict attention to fields for which  $1 + 1 \neq 0$ . Thus the phrase 'an arbitrary field' will mean 'an arbitrary field of characteristic  $\neq 2$ '. The proofs of many results make essential use of this restriction on the

characteristic. For any field  $F$ , we will denote by  $F^\times$  the multiplicative group of all nonzero elements of  $F$ . The squares in  $F^\times$  form a subgroup  $F^{\times 2}$  and any coset of this subgroup is called a *square class*.

Let  $V$  be a finite-dimensional vector space over such a field  $F$ . We say that  $V$  is a *quadratic space* if with each ordered pair  $u, v$  of elements of  $V$  there is associated an element  $(u, v)$  of  $F$  such that

- (i)  $(u_1 + u_2, v) = (u_1, v) + (u_2, v)$  for all  $u_1, u_2, v \in V$ ;
- (ii)  $(\alpha u, v) = \alpha(u, v)$  for every  $\alpha \in F$  and all  $u, v \in V$ ;
- (iii)  $(u, v) = (v, u)$  for all  $u, v \in V$ .

It follows that

- (i)'  $(u, v_1 + v_2) = (u, v_1) + (u, v_2)$  for all  $u, v_1, v_2 \in V$ ;
- (ii)'  $(u, \alpha v) = \alpha(u, v)$  for every  $\alpha \in F$  and all  $u, v \in V$ .

Let  $e_1, \dots, e_n$  be a basis for the vector space  $V$ . Then any  $u, v \in V$  can be uniquely expressed in the form

$$u = \sum_{j=1}^n \xi_j e_j, \quad v = \sum_{j=1}^n \eta_j e_j,$$

where  $\xi_j, \eta_j \in F$  ( $j = 1, \dots, n$ ), and

$$(u, v) = \sum_{j,k=1}^n \alpha_{jk} \xi_j \eta_k,$$

where  $\alpha_{jk} = (e_j, e_k) = \alpha_{kj}$ . Thus

$$(u, u) = \sum_{j,k=1}^n \alpha_{jk} \xi_j \xi_k$$

is a *quadratic form* with coefficients in  $F$ . The quadratic space is completely determined by the quadratic form, since

$$(u, v) = \{(u + v, u + v) - (u, u) - (v, v)\}/2. \quad (1)$$

Conversely, for a given basis  $e_1, \dots, e_n$  of  $V$ , any  $n \times n$  symmetric matrix  $A = (\alpha_{jk})$  with elements from  $F$ , or the associated quadratic form  $f(x) = x^t A x$ , may be used in this way to give  $V$  the structure of a quadratic space.

Let  $e'_1, \dots, e'_n$  be any other basis for  $V$ . Then

$$e_i = \sum_{j=1}^n \tau_{ji} e'_j,$$

where  $T = (\tau_{ji})$  is an invertible  $n \times n$  matrix with elements from  $F$ . Conversely, any such matrix  $T$  defines in this way a new basis  $e'_1, \dots, e'_n$ . Since

$$(e_i, e_k) = \sum_{j,h=1}^n \tau_{ji} \beta_{jh} \tau_{hk},$$

where  $\beta_{jh} = (e'_j, e'_h)$ , the matrix  $B = (\beta_{jh})$  is symmetric and

$$A = T^t B T. \quad (2)$$

Two symmetric matrices  $A, B$  with elements from  $F$  are said to be *congruent* if (2) holds for some invertible matrix  $T$  with elements from  $F$ . Thus congruence of symmetric matrices corresponds to a change of basis in the quadratic space. Evidently congruence is an equivalence relation, i.e. it is reflexive, symmetric and transitive. Two quadratic forms are said to be *equivalent over  $F$*  if their coefficient matrices are congruent. Equivalence over  $F$  of the quadratic forms  $f$  and  $g$  will be denoted by  $f \sim_F g$  or simply  $f \sim g$ .

It follows from (2) that

$$\det A = (\det T)^2 \det B.$$

Thus, although  $\det A$  is not uniquely determined by the quadratic space, if it is nonzero, its *square class* is uniquely determined. By abuse of language, we will call any representative of this square class the *determinant* of the quadratic space  $V$  and denote it by  $\det V$ .

Although quadratic spaces are better adapted for proving theorems, quadratic forms and symmetric matrices are useful for computational purposes. Thus a familiarity with both languages is desirable. However, we do not feel obliged to give two versions of each definition or result, and a version in one language may later be used in the other without explicit comment.

A vector  $v$  is said to be *orthogonal* to a vector  $u$  if  $(u, v) = 0$ . Then also  $u$  is orthogonal to  $v$ . The *orthogonal complement*  $U^\perp$  of a subspace  $U$  of  $V$  is defined to be the set of all  $v \in V$  such that  $(u, v) = 0$  for every  $u \in U$ . Evidently  $U^\perp$  is again a subspace. A subspace  $U$  will be said to be *non-singular* if  $U \cap U^\perp = \{0\}$ .

The whole space  $V$  is itself non-singular if and only if  $V^\perp = \{0\}$ . Thus  $V$  is non-singular if and only if some, and hence every, symmetric matrix describing it is non-singular, i.e. if and only if  $\det V \neq 0$ .

We say that a quadratic space  $V$  is the *orthogonal sum* of two subspaces  $V_1$  and  $V_2$ , and we write  $V = V_1 \perp V_2$ , if  $V = V_1 + V_2$ ,  $V_1 \cap V_2 = \{0\}$  and  $(v_1, v_2) = 0$  for all  $v_1 \in V_1, v_2 \in V_2$ .

If  $A_1$  is a coefficient matrix for  $V_1$  and  $A_2$  a coefficient matrix for  $V_2$ , then

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$$

is a coefficient matrix for  $V = V_1 \perp V_2$ . Thus  $\det V = (\det V_1)(\det V_2)$ . Evidently  $V$  is non-singular if and only if both  $V_1$  and  $V_2$  are non-singular.

If  $W$  is any subspace supplementary to the orthogonal complement  $V^\perp$  of the whole space  $V$ , then  $V = V^\perp \perp W$  and  $W$  is non-singular. Many problems for arbitrary quadratic spaces may be reduced in this way to non-singular quadratic spaces.

**Proposition 1** *If a quadratic space  $V$  contains a vector  $u$  such that  $(u, u) \neq 0$ , then*

$$V = U \perp U^\perp,$$

where  $U = \langle u \rangle$  is the one-dimensional subspace spanned by  $u$ .

*Proof* For any vector  $v \in V$ , put  $v' = v - \alpha u$ , where  $\alpha = (v, u)/(u, u)$ . Then  $(v', u) = 0$  and hence  $v' \in U^\perp$ . Since  $U \cap U^\perp = \{0\}$ , the result follows.  $\square$

A vector space basis  $u_1, \dots, u_n$  of a quadratic space  $V$  is said to be an *orthogonal basis* if  $(u_j, u_k) = 0$  whenever  $j \neq k$ .

**Proposition 2** *Any quadratic space  $V$  has an orthogonal basis.*

*Proof* If  $V$  has dimension 1, there is nothing to prove. Suppose  $V$  has dimension  $n > 1$  and the result holds for quadratic spaces of lower dimension. If  $(v, v) = 0$  for all  $v \in V$ , then any basis is an orthogonal basis, by (1). Hence we may assume that  $V$  contains a vector  $u_1$  such that  $(u_1, u_1) \neq 0$ . If  $U_1$  is the 1-dimensional subspace spanned by  $u_1$  then, by Proposition 1,

$$V = U_1 \perp U_1^\perp.$$

By the induction hypothesis  $U_1^\perp$  has an orthogonal basis  $u_2, \dots, u_n$ , and  $u_1, u_2, \dots, u_n$  is then an orthogonal basis for  $V$ .  $\square$

Proposition 2 says that any symmetric matrix  $A$  is congruent to a diagonal matrix, or that the corresponding quadratic form  $f$  is equivalent over  $F$  to a diagonal form  $\delta_1 \xi_1^2 + \dots + \delta_n \xi_n^2$ . Evidently  $\det f = \delta_1 \cdots \delta_n$  and  $f$  is non-singular if and only if  $\delta_j \neq 0$  ( $1 \leq j \leq n$ ). If  $A \neq 0$  then, by Propositions 1 and 2, we can take  $\delta_1$  to be any element of  $F^\times$  which is represented by  $f$ .

Here  $\gamma \in F^\times$  is said to be *represented* by a quadratic space  $V$  over the field  $F$  if there exists a vector  $v \in V$  such that  $(v, v) = \gamma$ .

As an application of Proposition 2 we prove

**Proposition 3** *If  $U$  is a non-singular subspace of the quadratic space  $V$ , then  $V = U \perp U^\perp$ .*

*Proof* Let  $u_1, \dots, u_m$  be an orthogonal basis for  $U$ . Then  $(u_j, u_j) \neq 0$  ( $1 \leq j \leq m$ ), since  $U$  is non-singular. For any vector  $v \in V$ , let  $u = \alpha_1 u_1 + \dots + \alpha_m u_m$ , where  $\alpha_j = (v, u_j)/(u_j, u_j)$  for each  $j$ . Then  $u \in U$  and  $(u, u_j) = (v, u_j)$  ( $1 \leq j \leq m$ ). Hence  $v - u \in U^\perp$ . Since  $U \cap U^\perp = \{0\}$ , the result follows.  $\square$

It may be noted that if  $U$  is a non-singular subspace and  $V = U \perp W$  for some subspace  $W$ , then necessarily  $W = U^\perp$ . For it is obvious that  $W \subseteq U^\perp$  and  $\dim W = \dim V - \dim U = \dim U^\perp$ , by Proposition 3.

**Proposition 4** *Let  $V$  be a non-singular quadratic space. If  $v_1, \dots, v_m$  are linearly independent vectors in  $V$  then, for any  $\eta_1, \dots, \eta_m \in F$ , there exists a vector  $v \in V$  such that  $(v_j, v) = \eta_j$  ( $1 \leq j \leq m$ ).*

Moreover, if  $U$  is any subspace of  $V$ , then

- (i)  $\dim U + \dim U^\perp = \dim V$ ;
- (ii)  $U^{\perp\perp} = U$ ;
- (iii)  $U^\perp$  is non-singular if and only if  $U$  is non-singular.

*Proof* There exist vectors  $v_{m+1}, \dots, v_n \in V$  such that  $v_1, \dots, v_n$  form a basis for  $V$ . If we put  $\alpha_{jk} = (v_j, v_k)$  then, since  $V$  is non-singular, the  $n \times n$  symmetric matrix  $A = (\alpha_{jk})$  is non-singular. Hence, for any  $\eta_1, \dots, \eta_n \in F$ , there exist unique  $\xi_1, \dots, \xi_n \in F$  such that  $v = \xi_1 v_1 + \dots + \xi_n v_n$  satisfies

$$(v_1, v) = \eta_1, \dots, (v_n, v) = \eta_n.$$

This proves the first part of the proposition.

By taking  $U = \langle v_1, \dots, v_m \rangle$  and  $\eta_1 = \dots = \eta_m = 0$ , we see that  $\dim U^\perp = n - m$ . Replacing  $U$  by  $U^\perp$ , we obtain  $\dim U^{\perp\perp} = \dim U$ . Since it is obvious that  $U \subseteq U^{\perp\perp}$ , this implies  $U = U^{\perp\perp}$ . Since  $U$  non-singular means  $U \cap U^\perp = \{0\}$ , (iii) follows at once from (ii).  $\square$

We now introduce some further definitions. A vector  $u$  is said to be *isotropic* if  $u \neq 0$  and  $(u, u) = 0$ . A subspace  $U$  of  $V$  is said to be *isotropic* if it contains an isotropic vector and *anisotropic* otherwise. A subspace  $U$  of  $V$  is said to be *totally isotropic* if every nonzero vector in  $U$  is isotropic, i.e. if  $U \subseteq U^\perp$ . According to these definitions, the trivial subspace  $\{0\}$  is both anisotropic and totally isotropic.

A quadratic space  $V$  over a field  $F$  is said to be *universal* if it represents every  $\gamma \in F^\times$ , i.e. if for each  $\gamma \in F^\times$  there is a vector  $v \in V$  such that  $(v, v) = \gamma$ .

**Proposition 5** *If a non-singular quadratic space  $V$  is isotropic, then it is universal.*

*Proof* Since  $V$  is isotropic, it contains a vector  $u \neq 0$  such that  $(u, u) = 0$ . Since  $V$  is non-singular, it contains a vector  $w$  such that  $(u, w) \neq 0$ . Then  $w$  is linearly independent of  $u$  and by replacing  $w$  by a scalar multiple we may assume  $(u, w) = 1$ . If  $v = au + w$ , then  $(v, v) = \gamma$  for  $a = \{\gamma - (w, w)\}/2$ .  $\square$

On the other hand, a non-singular universal quadratic space need not be isotropic. As an example, take  $F$  to be the finite field with three elements and  $V$  the 2-dimensional quadratic space corresponding to the quadratic form  $\xi_1^2 + \xi_2^2$ .

**Proposition 6** *A non-singular quadratic form  $f(\xi_1, \dots, \xi_n)$  with coefficients from a field  $F$  represents  $\gamma \in F^\times$  if and only if the quadratic form*

$$g(\xi_0, \xi_1, \dots, \xi_n) = -\gamma \xi_0^2 + f(\xi_1, \dots, \xi_n)$$

*is isotropic.*

*Proof* Obviously if  $f(x_1, \dots, x_n) = \gamma$  and  $x_0 = 1$ , then  $g(x_0, x_1, \dots, x_n) = 0$ . Suppose on the other hand that  $g(x_0, x_1, \dots, x_n) = 0$  for some  $x_j \in F$ , not all zero. If  $x_0 \neq 0$ , then  $f$  certainly represents  $\gamma$ . If  $x_0 = 0$ , then  $f$  is isotropic and hence, by Proposition 5, it still represents  $\gamma$ .  $\square$

**Proposition 7** *Let  $V$  be a non-singular isotropic quadratic space. If  $V = U \perp W$ , then there exists  $\gamma \in F^\times$  such that, for some  $u \in U$  and  $w \in W$ ,*

$$(u, u) = \gamma, \quad (w, w) = -\gamma.$$

*Proof* Since  $V$  is non-singular, so also are  $U$  and  $W$ , and since  $V$  contains an isotropic vector  $v'$ , there exist  $u' \in U$ ,  $w' \in W$ , not both zero, such that

$$(u', u') = -(w', w').$$

If this common value is nonzero, we are finished. Otherwise either  $U$  or  $W$  is isotropic. Without loss of generality, suppose  $U$  is isotropic. Since  $W$  is non-singular, it contains a vector  $w$  such that  $(w, w) \neq 0$ , and  $U$  contains a vector  $u$  such that  $(u, u) = -(w, w)$ , by Proposition 5.  $\square$

We now show that the totally isotropic subspaces of a quadratic space are important for an understanding of its structure, even though they are themselves trivial as quadratic spaces.

**Proposition 8** *All maximal totally isotropic subspaces of a quadratic space have the same dimension.*

*Proof* Let  $U_1$  be a maximal totally isotropic subspace of the quadratic space  $V$ . Then  $U_1 \subseteq U_1^\perp$  and  $U_1^\perp \setminus U_1$  contains no isotropic vector. Since  $V^\perp \subseteq U_1^\perp$ , it follows that  $V^\perp \subseteq U_1$ . If  $V'$  is a subspace of  $V$  supplementary to  $V^\perp$ , then  $V'$  is non-singular and  $U_1 = V^\perp + U'_1$ , where  $U'_1 \subseteq V'$ . Since  $U'_1$  is a maximal totally isotropic subspace of  $V'$ , this shows that it is sufficient to establish the result when  $V$  itself is non-singular.

Let  $U_2$  be another maximal totally isotropic subspace of  $V$ . Put  $W = U_1 \cap U_2$  and let  $W_1, W_2$  be subspaces supplementary to  $W$  in  $U_1, U_2$  respectively. We are going to show that  $W_2 \cap W_1^\perp = \{0\}$ .

Let  $v \in W_2 \cap W_1^\perp$ . Since  $W_2 \subseteq U_2$ ,  $v$  is isotropic and  $v \in U_2^\perp \subseteq W^\perp$ . Hence  $v \in U_1^\perp$  and actually  $v \in U_1$ , since  $v$  is isotropic. Since  $W_2 \subseteq U_2$  this implies  $v \in W$ , and since  $W \cap W_2 = \{0\}$  this implies  $v = 0$ .

It follows that  $\dim W_2 + \dim W_1^\perp \leq \dim V$ . But, since  $V$  is now assumed non-singular,  $\dim W_1 = \dim V - \dim W_1^\perp$ , by Proposition 4. Hence  $\dim W_2 \leq \dim W_1$  and, for the same reason,  $\dim W_1 \leq \dim W_2$ . Thus  $\dim W_2 = \dim W_1$ , and hence  $\dim U_2 = \dim U_1$ .  $\square$

We define the *index*,  $\text{ind } V$ , of a quadratic space  $V$  to be the dimension of any maximal totally isotropic subspace. Thus  $V$  is anisotropic if and only if  $\text{ind } V = 0$ .

A field  $F$  is said to be *ordered* if it contains a subset  $P$  of *positive* elements, which is closed under addition and multiplication, such that  $F$  is the disjoint union of the sets  $\{0\}$ ,  $P$  and  $-P = \{-x : x \in P\}$ . The rational field  $\mathbb{Q}$  and the real field  $\mathbb{R}$  are ordered fields, with the usual interpretation of 'positive'. For quadratic spaces over an ordered field there are other useful notions of index.

A subspace  $U$  of a quadratic space  $V$  over an ordered field  $F$  is said to be *positive definite* if  $(u, u) > 0$  for all nonzero  $u \in U$  and *negative definite* if  $(u, u) < 0$  for all nonzero  $u \in U$ . Evidently positive definite and negative definite subspaces are anisotropic.

**Proposition 9** *All maximal positive definite subspaces of a quadratic space  $V$  over an ordered field  $F$  have the same dimension.*

*Proof* Let  $U_+$  be a maximal positive definite subspace of the quadratic space  $V$ . Since  $U_+$  is certainly non-singular, we have  $V = U_+ \perp W$ , where  $W = U_+^\perp$ , and since  $U_+$  is maximal,  $(w, w) \leq 0$  for all  $w \in W$ . Since  $U_+ \subseteq V$ , we have  $V^\perp \subseteq W$ . If  $U_-$  is a maximal negative definite subspace of  $W$ , then in the same way  $W = U_- \perp U_0$ , where  $U_0 = U_-^\perp \cap W$ . Evidently  $U_0$  is totally isotropic and  $U_0 \subseteq V^\perp$ . In fact  $U_0 = V^\perp$ , since  $U_- \cap V^\perp = \{0\}$ . Since  $(v, v) \geq 0$  for all  $v \in U_+ \perp V^\perp$ , it follows that  $U_-$  is a maximal negative definite subspace of  $V$ .

If  $U'_+$  is another maximal positive definite subspace of  $V$ , then  $U'_+ \cap W = \{0\}$  and hence

$$\dim U'_+ + \dim W = \dim(U'_+ + W) \leq \dim V.$$

Thus  $\dim U'_+ \leq \dim V - \dim W = \dim U_+$ . But  $U_+$  and  $U'_+$  can be interchanged.  $\square$

If  $V$  is a quadratic space over an ordered field  $F$ , we define the *positive index*  $\text{ind}^+ V$  to be the dimension of any maximal positive definite subspace. Similarly all maximal negative definite subspaces have the same dimension, which we will call the *negative index* of  $V$  and denote by  $\text{ind}^- V$ . The proof of Proposition 9 shows that

$$\text{ind}^+ V + \text{ind}^- V + \dim V^\perp = \dim V.$$

**Proposition 10** *Let  $F$  denote the real field  $\mathbb{R}$  or, more generally, an ordered field in which every positive element is a square. Then any non-singular quadratic form  $f$  in  $n$  variables with coefficients from  $F$  is equivalent over  $F$  to a quadratic form*

$$g = \xi_1^2 + \cdots + \xi_p^2 - \xi_{p+1}^2 - \cdots - \xi_n^2,$$

where  $p \in \{0, 1, \dots, n\}$  is uniquely determined by  $f$ . In fact,

$$\text{ind}^+ f = p, \text{ind}^- f = n - p, \text{ind} f = \min(p, n - p).$$

*Proof* By Proposition 2,  $f$  is equivalent over  $F$  to a diagonal form  $\delta_1 \eta_1^2 + \cdots + \delta_n \eta_n^2$ , where  $\delta_j \neq 0$  ( $1 \leq j \leq n$ ). We may choose the notation so that  $\delta_j > 0$  for  $j \leq p$  and  $\delta_j < 0$  for  $j > p$ . The change of variables  $\xi_j = \delta_j^{1/2} \eta_j$  ( $j \leq p$ ),  $\xi_j = (-\delta_j)^{1/2} \eta_j$  ( $j > p$ ) now brings  $f$  to the form  $g$ . Since the corresponding quadratic space has a  $p$ -dimensional maximal positive definite subspace,  $p = \text{ind}^+ f$  is uniquely determined. Similarly  $n - p = \text{ind}^- f$ , and the formula for  $\text{ind} f$  follows readily.  $\square$

It follows that, for quadratic spaces over a field of the type considered in Proposition 10, a subspace is anisotropic if and only if it is either positive definite or negative definite.

Proposition 10 completely solves the problem of equivalence for real quadratic forms. (The uniqueness of  $p$  is known as *Sylvester's law of inertia*.) It will now be shown that the problem of equivalence for quadratic forms over a finite field can also be completely solved.

**Lemma 11** *If  $V$  is a non-singular 2-dimensional quadratic space over a finite field  $\mathbb{F}_q$ , of (odd) cardinality  $q$ , then  $V$  is universal.*

*Proof* By choosing an orthogonal basis for  $V$  we are reduced to showing that if  $\alpha, \beta, \gamma \in \mathbb{F}_q^\times$ , then there exist  $\xi, \eta \in \mathbb{F}_q$  such that  $\alpha\xi^2 + \beta\eta^2 = \gamma$ . As  $\xi$  runs through  $\mathbb{F}_q$ ,  $\alpha\xi^2$  takes  $(q+1)/2 = 1 + (q-1)/2$  distinct values. Similarly, as  $\eta$  runs through  $\mathbb{F}_q$ ,  $\gamma - \beta\eta^2$  takes  $(q+1)/2$  distinct values. Since  $(q+1)/2 + (q+1)/2 > q$ , there exist  $\xi, \eta \in \mathbb{F}_q$  for which  $\alpha\xi^2$  and  $\gamma - \beta\eta^2$  take the same value.  $\square$

**Proposition 12** Any non-singular quadratic form  $f$  in  $n$  variables over a finite field  $\mathbb{F}_q$  is equivalent over  $\mathbb{F}_q$  to the quadratic form

$$\xi_1^2 + \cdots + \xi_{n-1}^2 + \delta\xi_n^2,$$

where  $\delta = \det f$  is the determinant of  $f$ .

There are exactly two equivalence classes of non-singular quadratic forms in  $n$  variables over  $\mathbb{F}_q$ , one consisting of those forms  $f$  whose determinant  $\det f$  is a square in  $\mathbb{F}_q^\times$ , and the other those for which  $\det f$  is not a square in  $\mathbb{F}_q^\times$ .

*Proof* Since the first statement of the proposition is trivial for  $n = 1$ , we assume that  $n > 1$  and it holds for all smaller values of  $n$ . It follows from Lemma 11 that  $f$  represents 1 and hence, by the remark after the proof of Proposition 2,  $f$  is equivalent over  $\mathbb{F}_q$  to a quadratic form  $\xi_1^2 + g(\xi_2, \dots, \xi_n)$ . Since  $f$  and  $g$  have the same determinant, the first statement of the proposition now follows from the induction hypothesis.

Since  $\mathbb{F}_q^\times$  contains  $(q-1)/2$  distinct squares, every element of  $\mathbb{F}_q^\times$  is either a square or a square times a fixed non-square. The second statement of the proposition now follows from the first.  $\square$

We now return to quadratic spaces over an arbitrary field. A 2-dimensional quadratic space is said to be a *hyperbolic plane* if it is non-singular and isotropic.

**Proposition 13** For a 2-dimensional quadratic space  $V$ , the following statements are equivalent:

- (i)  $V$  is a hyperbolic plane;
- (ii)  $V$  has a basis  $u_1, u_2$  such that  $(u_1, u_1) = (u_2, u_2) = 0, (u_1, u_2) = 1$ ;
- (iii)  $V$  has a basis  $v_1, v_2$  such that  $(v_1, v_1) = 1, (v_2, v_2) = -1, (v_1, v_2) = 0$ ;
- (iv)  $-\det V$  is a square in  $F^\times$ .

*Proof* Suppose first that  $V$  is a hyperbolic plane and let  $u_1$  be any isotropic vector in  $V$ . If  $v$  is any linearly independent vector, then  $(u_1, v) \neq 0$ , since  $V$  is non-singular. By replacing  $v$  by a scalar multiple we may assume that  $(u_1, v) = 1$ . If we put  $u_2 = v + \alpha u_1$ , where  $\alpha = -(v, v)/2$ , then

$$(u_2, u_2) = (v, v) + 2\alpha = 0, (u_1, u_2) = (u_1, v) = 1,$$

and  $u_1, u_2$  is a basis for  $V$ .

If  $u_1, u_2$  are isotropic vectors in  $V$  such that  $(u_1, u_2) = 1$ , then the vectors  $v_1 = u_1 + u_2/2$  and  $v_2 = u_1 - u_2/2$  satisfy (iii), and if  $v_1, v_2$  satisfy (iii) then  $\det V = -1$ .

Finally, if (iv) holds then  $V$  is certainly non-singular. Let  $w_1, w_2$  be an orthogonal basis for  $V$  and put  $\delta_j = (w_j, w_j)$  ( $j = 1, 2$ ). By hypothesis,  $\delta_1\delta_2 = -\gamma^2$ , where  $\gamma \in F^\times$ . Since  $\gamma w_1 + \delta_1 w_2$  is an isotropic vector, this proves that (iv) implies (i).  $\square$

**Proposition 14** *Let  $V$  be a non-singular quadratic space. If  $U$  is a totally isotropic subspace with basis  $u_1, \dots, u_m$ , then there exists a totally isotropic subspace  $U'$  with basis  $u'_1, \dots, u'_m$  such that*

$$(u_j, u'_k) = 1 \text{ or } 0 \text{ according as } j = k \text{ or } j \neq k.$$

Hence  $U \cap U' = \{0\}$  and

$$U + U' = H_1 \perp \dots \perp H_m,$$

where  $H_j$  is the hyperbolic plane with basis  $u_j, u'_j$  ( $1 \leq j \leq m$ ).

*Proof* Suppose first that  $m = 1$ . Since  $V$  is non-singular, there exists a vector  $v \in V$  such that  $(u_1, v) \neq 0$ . The subspace  $H_1$  spanned by  $u_1, v$  is a hyperbolic plane and hence, by Proposition 13, it contains a vector  $u'_1$  such that  $(u'_1, u'_1) = 0$ ,  $(u_1, u'_1) = 1$ . This proves the proposition for  $m = 1$ .

Suppose now that  $m > 1$  and the result holds for all smaller values of  $m$ . Let  $W$  be the totally isotropic subspace with basis  $u_2, \dots, u_m$ . By Proposition 4, there exists a vector  $v \in W^\perp$  such that  $(u_1, v) \neq 0$ . The subspace  $H_1$  spanned by  $u_1, v$  is a hyperbolic plane and hence it contains a vector  $u'_1$  such that  $(u'_1, u'_1) = 0$ ,  $(u_1, u'_1) = 1$ . Since  $H_1$  is non-singular,  $H_1^\perp$  is also non-singular and  $V = H_1 \perp H_1^\perp$ . Since  $W \subseteq H_1^\perp$ , the result now follows by applying the induction hypothesis to the subspace  $W$  of the quadratic space  $H_1^\perp$ .  $\square$

**Proposition 15** *Any quadratic space  $V$  can be represented as an orthogonal sum*

$$V = V^\perp \perp H_1 \perp \dots \perp H_m \perp V_0,$$

where  $H_1, \dots, H_m$  are hyperbolic planes and the subspace  $V_0$  is anisotropic.

*Proof* Let  $V_1$  be any subspace supplementary to  $V^\perp$ . Then  $V_1$  is non-singular, by the definition of  $V^\perp$ . If  $V_1$  is anisotropic, we can take  $m = 0$  and  $V_0 = V_1$ . Otherwise  $V_1$  contains an isotropic vector and hence also a hyperbolic plane  $H_1$ , by Proposition 14. By Proposition 3,

$$V_1 = H_1 \perp V_2,$$

where  $V_2 = H_1^\perp \cap V_1$  is non-singular. If  $V_2$  is anisotropic, we can take  $V_0 = V_2$ . Otherwise we repeat the process. After finitely many steps we must obtain a representation of the required form, possibly with  $V_0 = \{0\}$ .  $\square$

Let  $V$  and  $V'$  be quadratic spaces over the same field  $F$ . The quadratic spaces  $V, V'$  are said to be *isometric* if there exists a linear map  $\varphi : V \rightarrow V'$  which is an *isometry*, i.e. it is bijective and

$$(\varphi v, \varphi v) = (v, v) \quad \text{for all } v \in V.$$

By (1), this implies

$$(\varphi u, \varphi v) = (u, v) \quad \text{for all } u, v \in V.$$

The concept of isometry is only another way of looking at equivalence. For if  $\varphi : V \rightarrow V'$  is an isometry, then  $V$  and  $V'$  have the same dimension. If  $u_1, \dots, u_n$  is a basis for  $V$  and  $u'_1, \dots, u'_n$  a basis for  $V'$  then, since  $(u_j, u_k) = (\varphi u_j, \varphi u_k)$ , the isometry is completely determined by the change of basis in  $V'$  from  $\varphi u_1, \dots, \varphi u_n$  to  $u'_1, \dots, u'_n$ .

A particularly simple type of isometry is defined in the following way. Let  $V$  be a quadratic space and  $w$  a vector such that  $(w, w) \neq 0$ . The map  $\tau : V \rightarrow V$  defined by

$$\tau v = v - \{2(v, w)/(w, w)\}w$$

is obviously linear. If  $W$  is the non-singular one-dimensional subspace spanned by  $w$ , then  $V = W \perp W^\perp$ . Since  $\tau v = v$  if  $v \in W^\perp$  and  $\tau v = -v$  if  $v \in W$ , it follows that  $\tau$  is bijective. Writing  $\alpha = -2(v, w)/(w, w)$ , we have

$$(\tau v, \tau v) = (v, v) + 2\alpha(v, w) + \alpha^2(w, w) = (v, v).$$

Thus  $\tau$  is an isometry. Geometrically,  $\tau$  is a *reflection* in the hyperplane orthogonal to  $w$ . We will refer to  $\tau = \tau_w$  as the reflection corresponding to the non-isotropic vector  $w$ .

**Proposition 16** *If  $u, u'$  are vectors of a quadratic space  $V$  such that  $(u, u) = (u', u') \neq 0$ , then there exists an isometry  $\varphi : V \rightarrow V$  such that  $\varphi u = u'$ .*

*Proof* Since

$$(u + u', u + u') + (u - u', u - u') = 2(u, u) + 2(u', u') = 4(u, u),$$

at least one of the vectors  $u + u', u - u'$  is not isotropic. If  $u - u'$  is not isotropic, the reflection  $\tau$  corresponding to  $w = u - u'$  has the property  $\tau u = u'$ , since  $(u - u', u - u') = 2(u, u - u')$ . If  $u + u'$  is not isotropic, the reflection  $\tau$  corresponding to  $w = u + u'$  has the property  $\tau u = -u'$ . Since  $u'$  is not isotropic, the corresponding reflection  $\sigma$  maps  $u'$  onto  $-u'$ , and hence the isometry  $\sigma \tau$  maps  $u$  onto  $u'$ .  $\square$

The proof of Proposition 16 has the following interesting consequence:

**Proposition 17** *Any isometry  $\varphi : V \rightarrow V$  of a non-singular quadratic space  $V$  is a product of reflections.*

*Proof* Let  $u_1, \dots, u_n$  be an orthogonal basis for  $V$ . By Proposition 16 and its proof, there exists an isometry  $\psi$ , which is either a reflection or a product of two reflections, such that  $\psi u_1 = \varphi u_1$ . If  $U$  is the subspace with basis  $u_1$  and  $W$  the subspace with basis  $u_2, \dots, u_n$ , then  $V = U \perp W$  and  $W = U^\perp$  is non-singular. Since the isometry  $\varphi_1 = \psi^{-1} \varphi$  fixes  $u_1$ , we have also  $\varphi_1 W = W$ . But if  $\sigma : W \rightarrow W$  is a reflection, the extension  $\tau : V \rightarrow V$  defined by  $\tau u = u$  if  $u \in U$ ,  $\tau w = \sigma w$  if  $w \in W$ , is also a reflection. By using induction on the dimension  $n$ , it follows that  $\varphi_1$  is a product of reflections, and hence so also is  $\varphi = \psi \varphi_1$ .  $\square$

By a more elaborate argument E. Cartan (1938) showed that any isometry of an  $n$ -dimensional non-singular quadratic space is a product of at most  $n$  reflections.

**Proposition 18** *Let  $V$  be a quadratic space with two orthogonal sum representations*

$$V = U \perp W = U' \perp W'.$$

*If there exists an isometry  $\phi : U \rightarrow U'$ , then there exists an isometry  $\psi : V \rightarrow V$  such that  $\psi u = \phi u$  for all  $u \in U$  and  $\psi W = W'$ . Thus if  $U$  is isometric to  $U'$ , then  $W$  is isometric to  $W'$ .*

*Proof* Let  $u_1, \dots, u_m$  and  $u_{m+1}, \dots, u_n$  be bases for  $U$  and  $W$  respectively. If  $u'_j = \phi u_j$  ( $1 \leq j \leq m$ ), then  $u'_1, \dots, u'_m$  is a basis for  $U'$ . Let  $u'_{m+1}, \dots, u'_n$  be a basis for  $W'$ . The symmetric matrices associated with the bases  $u_1, \dots, u_n$  and  $u'_1, \dots, u'_n$  of  $V$  have the form

$$\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}, \begin{pmatrix} A & 0 \\ 0 & C \end{pmatrix},$$

which we will write as  $A \oplus B$ ,  $A \oplus C$ . Thus the two matrices  $A \oplus B$ ,  $A \oplus C$  are congruent. It is enough to show that this implies that  $B$  and  $C$  are congruent. For suppose  $C = S'BS$  for some invertible matrix  $S = (\sigma_{ij})$ . If we define  $u''_{m+1}, \dots, u''_n$  by

$$u''_i = \sum_{j=m+1}^n \sigma_{ji} u'_j \quad (m+1 \leq i \leq n),$$

then  $(u''_j, u''_k) = (u_j, u_k)$  ( $m+1 \leq j, k \leq n$ ) and the linear map  $\psi : V \rightarrow V$  defined by

$$\psi u_j = u'_j \quad (1 \leq j \leq m), \quad \psi u_j = u''_j \quad (m+1 \leq j \leq n),$$

is the required isometry.

By taking the bases for  $U$ ,  $W$ ,  $W'$  to be orthogonal bases we are reduced to the case in which  $A, B, C$  are diagonal matrices. We may choose the notation so that  $A = \text{diag}[a_1, \dots, a_m]$ , where  $a_j \neq 0$  for  $j \leq r$  and  $a_j = 0$  for  $j > r$ . If  $a_1 \neq 0$ , i.e. if  $r > 0$ , and if we write  $A' = \text{diag}[a_2, \dots, a_m]$ , then it follows from Propositions 1 and 16 that the matrices  $A' \oplus B$  and  $A' \oplus C$  are congruent. Proceeding in this way, we are reduced to the case  $A = O$ .

Thus we now suppose  $A = O$ . We may assume  $B \neq O$ ,  $C \neq O$ , since otherwise the result is obvious. We may choose the notation also so that  $B = O_s \oplus B'$  and  $C = O_s \oplus C'$ , where  $B'$  is non-singular and  $0 \leq s < n - m$ . If  $T'(O_{m+s} \oplus C')T = O_{m+s} \oplus B'$ , where

$$T = \begin{pmatrix} T_1 & T_2 \\ T_3 & T_4 \end{pmatrix},$$

then  $T_4^t C' T_4 = B'$ . Since  $B'$  is non-singular, so also is  $T_4$  and thus  $B'$  and  $C'$  are congruent. It follows that  $B$  and  $C$  are also congruent.  $\square$

**Corollary 19** *If a non-singular subspace  $U$  of a quadratic space  $V$  is isometric to another subspace  $U'$ , then  $U^\perp$  is isometric to  $U'^\perp$ .*

*Proof* This follows at once from Proposition 18, since  $U'$  is also non-singular and

$$V = U \perp U^\perp = U' \perp U'^\perp. \quad \square$$

The first statement of Proposition 18 is known as *Witt's extension theorem* and the second statement as *Witt's cancellation theorem*. It was Corollary 19 which was actually proved by Witt (1937).

There is also another version of the extension theorem, stating that if  $\phi : U \rightarrow U'$  is an isometry between two subspaces  $U, U'$  of a *non-singular* quadratic space  $V$ , then there exists an isometry  $\psi : V \rightarrow V$  such that  $\psi u = \phi u$  for all  $u \in U$ . For non-singular  $U$  this has just been proved, and the singular case can be reduced to the non-singular by applying (several times, if necessary) the following lemma.

**Lemma 20** *Let  $V$  be a non-singular quadratic space. If  $U, U'$  are singular subspaces of  $V$  and if there exists an isometry  $\phi : U \rightarrow U'$ , then there exist subspaces  $\bar{U}, \bar{U}'$ , properly containing  $U, U'$  respectively and an isometry  $\bar{\phi} : \bar{U} \rightarrow \bar{U}'$  such that  $\bar{\phi}u = \phi u$  for all  $u \in U$ .*

*Proof* By hypothesis there exists a nonzero vector  $u_1 \in U \cap U^\perp$ . Then  $U$  has a basis  $u_1, \dots, u_m$  with  $u_1$  as first vector. By Proposition 4, there exists a vector  $w \in V$  such that

$$(u_1, w) = 1, (u_j, w) = 0 \quad \text{for } 1 < j \leq m.$$

Moreover we may assume that  $(w, w) = 0$ , by replacing  $w$  by  $w - \alpha u_1$ , with  $\alpha = (w, w)/2$ . If  $W$  is the 1-dimensional subspace spanned by  $w$ , then  $U \cap W = \{0\}$  and  $\bar{U} = U + W$  contains  $U$  properly.

The same construction can be applied to  $U'$ , with the basis  $\phi u_1, \dots, \phi u_m$ , to obtain an isotropic vector  $w'$  and a subspace  $\bar{U}' = U' + W'$ . The linear map  $\bar{\phi} : \bar{U} \rightarrow \bar{U}'$  defined by

$$\bar{\phi}u_j = \phi u_j \quad (1 \leq j \leq m), \quad \bar{\phi}w = w',$$

is easily seen to have the required properties.  $\square$

As an application of Proposition 18, we will consider the uniqueness of the representation obtained in Proposition 15.

**Proposition 21** *Suppose the quadratic space  $V$  can be represented as an orthogonal sum*

$$V = U \perp H \perp V_0,$$

*where  $U$  is totally isotropic,  $H$  is the orthogonal sum of  $m$  hyperbolic planes, and the subspace  $V_0$  is anisotropic.*

*Then  $U = V^\perp$ ,  $m = \text{ind } V - \dim V^\perp$ , and  $V_0$  is uniquely determined up to an isometry.*

*Proof* Since  $H$  and  $V_0$  are non-singular, so also is  $W = H \perp V_0$ . Hence, by the remark after the proof of Proposition 3,  $U = W^\perp$ . Since  $U \subseteq U^\perp$ , it follows that  $U \subseteq V^\perp$ . In fact  $U = V^\perp$ , since  $W \cap V^\perp = \{0\}$ .

The subspace  $H$  has two  $m$ -dimensional totally isotropic subspaces  $U_1, U'_1$  such that

$$H = U_1 + U'_1, \quad U_1 \cap U'_1 = \{0\}.$$

Evidently  $V_1 := V^\perp + U_1$  is a totally isotropic subspace of  $V$ . In fact  $V_1$  is maximal, since any isotropic vector in  $U'_1 \perp V_0$  is contained in  $U'_1$ . Thus  $m = \text{ind } V - \dim V^\perp$  is uniquely determined and  $H$  is uniquely determined up to an isometry. If also

$$V = V^\perp \perp H' \perp V'_0,$$

where  $H'$  is the orthogonal sum of  $m$  hyperbolic planes and  $V'_0$  is anisotropic then, by Proposition 18,  $V_0$  is isometric to  $V'_0$ .  $\square$

Proposition 21 reduces the problem of equivalence for quadratic forms over an arbitrary field to the case of anisotropic forms. As we will see, this can still be a difficult problem, even for the field of rational numbers.

Two quadratic spaces  $V, V'$  over the same field  $F$  may be said to be *Witt-equivalent*, in symbols  $V \approx V'$ , if their anisotropic components  $V_0, V'_0$  are isometric. This is certainly an equivalence relation. The cancellation law makes it possible to define various algebraic operations on the set  $\mathcal{W}(F)$  of all quadratic spaces over the field  $F$ , with equality replaced by Witt-equivalence. If we define  $-V$  to be the quadratic space with the same underlying vector space as  $V$  but with  $(v_1, v_2)$  replaced by  $-(v_1, v_2)$ , then

$$V \perp (-V) \approx \{0\}.$$

If we define the *sum* of two quadratic spaces  $V$  and  $W$  to be  $V \perp W$ , then

$$V \approx V', \quad W \approx W' \Rightarrow V \perp W \approx V' \perp W'.$$

Similarly, if we define the *product* of  $V$  and  $W$  to be the tensor product  $V \otimes W$  of the underlying vector spaces with the quadratic space structure defined by

$$(\{v_1, w_1\}, \{v_2, w_2\}) = (v_1, v_2)(w_1, w_2),$$

then

$$V \approx V', \quad W \approx W' \Rightarrow V \otimes W \approx V' \otimes W'.$$

It is readily seen that in this way  $\mathcal{W}(F)$  acquires the structure of a commutative ring, the *Witt ring* of the field  $F$ .

## 2 The Hilbert Symbol

Again let  $F$  be any field of characteristic  $\neq 2$  and  $F^\times$  the multiplicative group of all nonzero elements of  $F$ . We define the *Hilbert symbol*  $(a, b)_F$ , where  $a, b \in F^\times$ , by

$$\begin{aligned} (a, b)_F &= 1 \text{ if there exist } x, y \in F \text{ such that } ax^2 + by^2 = 1, \\ &= -1 \text{ otherwise.} \end{aligned}$$

By Proposition 6,  $(a, b)_F = 1$  if and only if the ternary quadratic form  $a\xi^2 + b\eta^2 - \zeta^2$  is isotropic.

The following lemma shows that the Hilbert symbol can also be defined in an asymmetric way:

**Lemma 22** *For any field  $F$  and any  $a, b \in F^\times$ ,  $(a, b)_F = 1$  if and only if the binary quadratic form  $f_a = \xi^2 - a\eta^2$  represents  $b$ . Moreover, for any  $a \in F^\times$ , the set  $G_a$  of all  $b \in F^\times$  which are represented by  $f_a$  is a subgroup of  $F^\times$ .*

*Proof* Suppose first that  $ax^2 + by^2 = 1$  for some  $x, y \in F$ . If  $a$  is a square, the quadratic form  $f_a$  is isotropic and hence universal. If  $a$  is not a square, then  $y \neq 0$  and  $(y^{-1})^2 - a(xy^{-1})^2 = b$ .

Suppose next that  $u^2 - av^2 = b$  for some  $u, v \in F$ . If  $-ba^{-1}$  is a square, the quadratic form  $a\xi^2 + b\eta^2$  is isotropic and hence universal. If  $-ba^{-1}$  is not a square, then  $u \neq 0$  and  $a(vu^{-1})^2 + b(u^{-1})^2 = 1$ .

It is obvious that if  $b \in G_a$ , then also  $b^{-1} \in G_a$ , and it is easily verified that if

$$\zeta_1 = \xi_1\eta_1 + a\xi_2\eta_2, \quad \zeta_2 = \xi_1\eta_2 + \xi_2\eta_1,$$

then

$$\zeta_1^2 - a\zeta_2^2 = (\xi_1^2 - a\xi_2^2)(\eta_1^2 - a\eta_2^2).$$

(In fact this is just Brahmagupta's identity, already encountered in §4 of Chapter IV.) It follows that  $G_a$  is a subgroup of  $F^\times$ .  $\square$

**Proposition 23** *For any field  $F$ , the Hilbert symbol has the following properties:*

- (i)  $(a, b)_F = (b, a)_F$ ,
- (ii)  $(a, bc^2)_F = (a, b)_F$  for any  $c \in F^\times$ ,
- (iii)  $(a, 1)_F = 1$ ,
- (iv)  $(a, -ab)_F = (a, b)_F$ ,
- (v) if  $(a, b)_F = 1$ , then  $(a, bc)_F = (a, c)_F$  for any  $c \in F^\times$ .

*Proof* The first three properties follow immediately from the definition. The fourth property follows from Lemma 22. For, since  $G_a$  is a group and  $f_a$  represents  $-a$ ,  $f_a$  represents  $-ab$  if and only if it represents  $b$ . The proof of (v) is similar: if  $f_a$  represents  $b$ , then it represents  $bc$  if and only if it represents  $c$ .  $\square$

The Hilbert symbol will now be evaluated for the real field  $\mathbb{R} = \mathbb{Q}_\infty$  and the  $p$ -adic fields  $\mathbb{Q}_p$  studied in Chapter VI. In these cases it will be denoted simply by  $(a, b)_\infty$ , resp.  $(a, b)_p$ . For the real field, we obtain at once from the definition of the Hilbert symbol

**Proposition 24** *Let  $a, b \in \mathbb{R}^\times$ . Then  $(a, b)_\infty = -1$  if and only if both  $a < 0$  and  $b < 0$ .*

To evaluate  $(a, b)_p$ , we first note that we can write  $a = p^\alpha a'$ ,  $b = p^\beta b'$ , where  $\alpha, \beta \in \mathbb{Z}$  and  $|a'|_p = |b'|_p = 1$ . It follows from (i), (ii) of Proposition 23 that we may assume  $\alpha, \beta \in \{0, 1\}$ . Furthermore, by (ii), (iv) of Proposition 23 we may assume that  $\alpha$  and  $\beta$  are not both 1. Thus we are reduced to the case where  $a$  is a  $p$ -adic unit and either  $b$  is a  $p$ -adic unit or  $b = pb'$ , where  $b'$  is a  $p$ -adic unit. To evaluate  $(a, b)_p$  under these assumptions we will use the conditions for a  $p$ -adic unit to be a square which were derived in Chapter VI. It is convenient to treat the case  $p = 2$  separately.

**Proposition 25** *Let  $p$  be an odd prime and  $a, b \in \mathbb{Q}_p$  with  $|a|_p = |b|_p = 1$ . Then*

- (i)  $(a, b)_p = 1$ ,
- (ii)  $(a, pb)_p = 1$  if and only if  $a = c^2$  for some  $c \in \mathbb{Q}_p$ .

*In particular, for any integers  $a, b$  not divisible by  $p$ ,  $(a, b)_p = 1$  and  $(a, pb)_p = (a/p)$ , where  $(a/p)$  is the Legendre symbol.*

*Proof* Let  $S \subseteq \mathbb{Z}_p$  be a set of representatives, with  $0 \in S$ , of the finite residue field  $\mathbb{F}_p = \mathbb{Z}_p/p\mathbb{Z}_p$ . There exist non-zero  $a_0, b_0 \in S$  such that

$$|a - a_0|_p < 1, |b - b_0|_p < 1.$$

But Lemma 11 implies that there exist  $x_0, y_0 \in S$  such that

$$|a_0x_0^2 + b_0y_0^2 - 1|_p < 1.$$

Since  $|x_0|_p \leq 1, |y_0|_p \leq 1$ , it follows that

$$|ax_0^2 + by_0^2 - 1|_p < 1.$$

Hence, by Proposition VI.16,  $ax_0^2 + by_0^2 = z^2$  for some  $z \in \mathbb{Q}_p$ . Since  $z \neq 0$ , this implies  $(a, b)_p = 1$ . This proves (i).

If  $a = c^2$  for some  $c \in \mathbb{Q}_p$ , then  $(a, pb)_p = 1$ , by Proposition 23. Conversely, suppose there exist  $x, y \in \mathbb{Q}_p$  such that  $ax^2 + pby^2 = 1$ . Then  $|ax^2|_p \neq |pby^2|_p$ , since  $|a|_p = |b|_p = 1$ . It follows that  $|x|_p = 1, |y|_p \leq 1$ . Thus  $|ax^2 - 1|_p < 1$  and hence  $ax^2 = z^2$  for some  $z \in \mathbb{Q}_p^\times$ . This proves (ii).

The special case where  $a$  and  $b$  are integers now follows from Corollary VI.17.  $\square$

**Corollary 26** *If  $p$  is an odd prime and if  $a, b, c \in \mathbb{Q}_p$  are  $p$ -adic units, then the quadratic form  $a\xi^2 + b\eta^2 + c\zeta^2$  is isotropic.*

*Proof* In fact, the quadratic form  $-c^{-1}a\xi^2 - c^{-1}b\eta^2 - \zeta^2$  is isotropic, since  $(-c^{-1}a, -c^{-1}b)_p = 1$ , by Proposition 25.  $\square$

**Proposition 27** *Let  $a, b \in \mathbb{Q}_2$  with  $|a|_2 = |b|_2 = 1$ . Then*

- (i)  $(a, b)_2 = 1$  if and only if at least one of  $a, b, a - 4, b - 4$  is a square in  $\mathbb{Q}_2$ ;
- (ii)  $(a, 2b)_2 = 1$  if and only if either  $a$  or  $a + 2b$  is a square in  $\mathbb{Q}_2$ .

*In particular, for any odd integers  $a, b$ ,  $(a, b)_2 = 1$  if and only if  $a \equiv 1$  or  $b \equiv 1 \pmod{4}$ , and  $(a, 2b)_2 = 1$  if and only if  $a \equiv 1$  or  $a + 2b \equiv 1 \pmod{8}$ .*

*Proof* Suppose there exist  $x, y \in \mathbb{Q}_2$  such that  $ax^2 + by^2 = 1$  and assume, for example, that  $|x|_2 \geq |y|_2$ . Then  $|x|_2 \geq 1$  and  $|x|_2 = 2^\alpha$ , where  $\alpha \geq 0$ . By Corollary VI.14,

$$x = 2^\alpha(x_0 + 4x'), \quad y = 2^\alpha(y_0 + 4y'),$$

where  $x_0 \in \{1, 3\}$ ,  $y_0 \in \{0, 1, 2, 3\}$  and  $x', y' \in \mathbb{Z}_2$ . If  $a$  and  $b$  are not squares in  $\mathbb{Q}_2$  then, by Proposition VI.16,  $|a - 1|_2 > 2^{-3}$  and  $|b - 1|_2 > 2^{-3}$ . Thus

$$a = a_0 + 8a', \quad b = b_0 + 8b',$$

where  $a_0, b_0 \in \{3, 5, 7\}$  and  $a', b' \in \mathbb{Z}_2$ . Hence

$$1 = ax^2 + by^2 = 2^{2a}(a_0 + b_0y_0^2 + 8z'),$$

where  $z' \in \mathbb{Z}_2$ . Since  $a_0, b_0$  are odd and  $y_0^2 \equiv 0, 1$  or  $4 \pmod{8}$ , we must have  $a = 0$ ,  $y_0^2 \equiv 4 \pmod{8}$  and  $a_0 = 5$ . Thus, by Proposition VI.16 again,  $a - 4$  is a square in  $\mathbb{Q}_2$ . This proves that the condition in (i) is necessary.

If  $a$  is a square in  $\mathbb{Q}_2$ , then certainly  $(a, b)_2 = 1$ . If  $a - 4$  is a square, then  $a = 5 + 8a'$ , where  $a' \in \mathbb{Z}_2$ , and  $a + 4b = 1 + 8c'$ , where  $c' \in \mathbb{Z}_2$ . Hence  $a + 4b$  is a square in  $\mathbb{Q}_2$  and the quadratic form  $a\xi^2 + b\eta^2$  represents 1. This proves that the condition in (i) is sufficient.

Suppose next that there exist  $x, y \in \mathbb{Q}_2$  such that  $ax^2 + 2by^2 = 1$ . By the same argument as for odd  $p$  in Proposition 25, we must have  $|x|_2 = 1$ ,  $|y|_2 \leq 1$ . Thus  $x = x_0 + 4x'$ ,  $y = y_0 + 4y'$ , where  $x_0 \in \{1, 3\}$ ,  $y_0 \in \{0, 1, 2, 3\}$  and  $x', y' \in \mathbb{Z}_2$ . Writing  $a = a_0 + 8a'$ ,  $b = b_0 + 8b'$ , where  $a_0, b_0 \in \{1, 3, 5, 7\}$  and  $a', b' \in \mathbb{Z}_2$ , we obtain  $a_0x_0^2 + 2b_0y_0^2 \equiv 1 \pmod{8}$ . Since  $2y_0^2 \equiv 0$  or  $2 \pmod{8}$ , this implies either  $a_0 \equiv 1$  or  $a_0 + 2b_0 \equiv 1 \pmod{8}$ . Hence either  $a$  or  $a + 2b$  is a square in  $\mathbb{Q}_2$ . It is obvious that, conversely,  $(a, 2b)_2 = 1$  if either  $a$  or  $a + 2b$  is a square in  $\mathbb{Q}_2$ .

The special case where  $a$  and  $b$  are integers again follows from Corollary VI.17.  $\square$

For  $F = \mathbb{R}$ , the factor group  $F^\times/F^{\times 2}$  is of order 2, with 1 and  $-1$  as representatives of the two square classes. For  $F = \mathbb{Q}_p$ , with  $p$  odd, it follows from Corollary VI.17 that the factor group  $F^\times/F^{\times 2}$  is of order 4. Moreover, if  $r$  is an integer such that  $(r/p) = -1$ , then  $1, r, p, rp$  are representatives of the four square classes. Similarly for  $F = \mathbb{Q}_2$ , the factor group  $F^\times/F^{\times 2}$  is of order 8 and  $1, 3, 5, 7, 2, 6, 10, 14$  are representatives of the eight square classes. The Hilbert symbol  $(a, b)_F$  for these representatives, and hence for all  $a, b \in F^\times$ , may be determined directly from Propositions 24, 25 and 27. The values obtained are listed in Table 1, where  $\varepsilon = (-1/p)$  and thus  $\varepsilon = \pm 1$  according as  $p \equiv \pm 1 \pmod{4}$ .

It will be observed that each of the three symmetric matrices in Table 1 is a Hadamard matrix! In particular, in each row after the first row of  $+$ 's there are equally many  $+$  and  $-$  signs. This property turns out to be of basic importance and prompts the following definition:

A field  $F$  is a *Hilbert field* if some  $a \in F^\times$  is not a square and if, for every such  $a$ , the subgroup  $G_a$  has index 2 in  $F^\times$ .

Thus the real field  $\mathbb{R} = \mathbb{Q}_\infty$  and the  $p$ -adic fields  $\mathbb{Q}_p$  are all Hilbert fields. We now show that in Hilbert fields further properties of the Hilbert symbol may be derived.

**Proposition 28** *A field  $F$  is a Hilbert field if and only if some  $a \in F^\times$  is not a square and the Hilbert symbol has the following additional properties:*

- (i) if  $(a, b)_F = 1$  for every  $b \in F^\times$ , then  $a$  is a square in  $F^\times$ ;
- (ii)  $(a, bc)_F = (a, b)_F(a, c)_F$  for all  $a, b, c \in F^\times$ .

*Proof* Let  $F$  be a Hilbert field. Then (i) holds, since  $G_a \neq F^\times$  if  $a$  is not a square. If  $(a, b)_F = 1$  or  $(a, c)_F = 1$ , then (ii) follows from Proposition 23(v). Suppose now that  $(a, b)_F = -1$  and  $(a, c)_F = -1$ . Then  $a$  is not a square and  $f_a$  does not represent  $b$  or  $c$ . Since  $F$  is a Hilbert field and  $b, c \notin G_a$ , it follows that  $bc \in G_a$ . Thus  $(a, bc)_F = 1$ . The converse is equally simple.  $\square$

**Table 1.** Values of the Hilbert symbol  $(a, b)_F$  for  $F = \mathbb{Q}_p$

$\mathbb{Q}_\infty = \mathbb{R}$			$\mathbb{Q}_p : p \text{ odd}$				
$a \backslash b$	1	-1	$a \backslash b$	1	$p$	$rp$	$r$
1	+	+	1	+	+	+	+
-1	+	-	$p$	+	$\varepsilon$	$-\varepsilon$	-
			$rp$	+	$-\varepsilon$	$\varepsilon$	-
			$r$	+	-	-	+

where  $r$  is a primitive root mod  $p$  and  $\varepsilon = (-1)^{(p-1)/2}$

$\mathbb{Q}_2$								
$a \backslash b$	1	3	6	2	14	10	5	7
1	+	+	+	+	+	+	+	+
3	+	-	+	-	+	-	+	-
6	+	+	-	-	+	+	-	-
2	+	-	-	+	+	-	-	+
14	+	+	+	+	-	-	-	-
10	+	-	+	-	-	+	-	+
5	+	+	-	-	-	-	+	+
7	+	-	-	+	-	+	+	-

The definition of a Hilbert field can be reformulated in terms of quadratic forms. If  $f$  is an anisotropic binary quadratic form with determinant  $d$ , then  $-d$  is not a square and  $f$  is equivalent to a diagonal form  $a(\zeta^2 + d\eta^2)$ . It follows that  $F$  is a Hilbert field if and only if there exists an anisotropic binary quadratic form and for each such form there is, apart from equivalent forms, exactly one other whose determinant is in the same square class. We are going to show that Hilbert fields can also be characterized by means of quadratic forms in 4 variables.

**Lemma 29** *Let  $F$  be an arbitrary field and  $a, b$  elements of  $F^\times$  with  $(a, b)_F = -1$ . Then the quadratic form*

$$f_{a,b} = \zeta_1^2 - a\zeta_2^2 - b(\zeta_3^2 - a\zeta_4^2)$$

*is anisotropic. Moreover, the set  $G_{a,b}$  of all elements of  $F^\times$  which are represented by  $f_{a,b}$  is a subgroup of  $F^\times$ .*

*Proof* Since  $(a, b)_F = -1$ ,  $a$  is not a square and hence the binary form  $f_a$  is anisotropic. If  $f_{a,b}$  were isotropic, some  $c \in F^\times$  would be represented by both  $f_a$  and  $bf_a$ . But then  $(a, c)_F = 1$  and  $(a, bc)_F = 1$ . Since  $(a, b)_F = -1$ , this contradicts Proposition 23.

Clearly if  $c \in G_{a,b}$ , then also  $c^{-1} \in G_{a,b}$ , and it is easily verified that if

$$\begin{aligned} \zeta_1 &= \zeta_1\eta_1 + a\zeta_2\eta_2 + b\zeta_3\eta_3 - ab\zeta_4\eta_4, & \zeta_2 &= \zeta_1\eta_2 + \zeta_2\eta_1 - b\zeta_3\eta_4 + b\zeta_4\eta_3, \\ \zeta_3 &= \zeta_1\eta_3 + \zeta_3\eta_1 + a\zeta_2\eta_4 - a\zeta_4\eta_2, & \zeta_4 &= \zeta_1\eta_4 + \zeta_4\eta_1 + \zeta_2\eta_3 - \zeta_3\eta_2, \end{aligned}$$

then

$$\zeta_1^2 - a\zeta_2^2 - b\zeta_3^2 + ab\zeta_4^2 = (\zeta_1^2 - a\zeta_2^2 - b\zeta_3^2 + ab\zeta_4^2)(\eta_1^2 - a\eta_2^2 - b\eta_3^2 + ab\eta_4^2).$$

It follows that  $G_{a,b}$  is a subgroup of  $F^\times$ .  $\square$

**Proposition 30** *A field  $F$  is a Hilbert field if and only if one of the following mutually exclusive conditions is satisfied:*

- (A)  *$F$  is an ordered field and every positive element of  $F$  is a square;*
- (B) *there exists, up to equivalence, one and only one anisotropic quaternary quadratic form over  $F$ .*

*Proof* Suppose first that the field  $F$  is of type (A). Then  $-1$  is not a square, since  $-1 + 1 = 0$  and any nonzero square is positive. By Proposition 10, any anisotropic binary quadratic form is equivalent over  $F$  to exactly one of the forms  $\zeta^2 + \eta^2$ ,  $-\zeta^2 - \eta^2$  and therefore  $F$  is a Hilbert field. Since the quadratic forms  $\zeta_1^2 + \zeta_2^2 + \zeta_3^2 + \zeta_4^2$  and  $-\zeta_1^2 - \zeta_2^2 - \zeta_3^2 - \zeta_4^2$  are anisotropic and inequivalent, the field  $F$  is not of type (B).

Suppose next that the field  $F$  is of type (B). The anisotropic quaternary quadratic form must be universal, since it is equivalent to any nonzero scalar multiple. Hence, for any  $a \in F^\times$  there exists an anisotropic diagonal form

$$-a\zeta_1^2 - b'\zeta_2^2 - c'\zeta_3^2 - d'\zeta_4^2,$$

where  $b', c', d' \in F^\times$ . In particular, for  $a = -1$ , this shows that not every element of  $F^\times$  is a square. The ternary quadratic form  $h = -b'\zeta_2^2 - c'\zeta_3^2 - d'\zeta_4^2$  is certainly anisotropic. If  $h$  does not represent 1, the quaternary quadratic form  $-\zeta_1^2 + h$  is also anisotropic and hence, by Witt's cancellation theorem,  $a$  must be a square. Consequently, if  $a \in F^\times$  is not a square, then there exists an anisotropic form

$$-a\zeta_1^2 + \zeta_2^2 - b\zeta_3^2 - c\zeta_4^2.$$

Thus for any  $a \in F^\times$  which is not a square, there exists  $b \in F^\times$  such that  $(a, b)_F = -1$ . If  $(a, b)_F = (a, b')_F = -1$  then, by Lemma 29, the forms

$$\zeta_1^2 - a\zeta_2^2 - b(\zeta_3^2 - a\zeta_4^2), \zeta_1^2 - a\zeta_2^2 - b'(\zeta_3^2 - a\zeta_4^2)$$

are anisotropic and thus equivalent. It follows from Witt's cancellation theorem that the binary forms  $b(\zeta_3^2 - a\zeta_4^2)$  and  $b'(\zeta_3^2 - a\zeta_4^2)$  are equivalent. Consequently  $\zeta_3^2 - a\zeta_4^2$  represents  $bb'$  and  $(a, bb')_F = 1$ . Thus  $G_a$  has index 2 in  $F^\times$  for any  $a \in F^\times$  which is not a square, and  $F$  is a Hilbert field.

Suppose now that  $F$  is a Hilbert field. Then there exists  $a \in F^\times$  which is not a square and, for any such  $a$ , there exists  $b \in F^\times$  such that  $(a, b)_F = -1$ . Consequently, by Lemma 29, the quaternary quadratic form  $f_{a,b}$  is anisotropic and represents 1. Conversely, any anisotropic quaternary quadratic form which represents 1 is equivalent to some form

$$g = \zeta_1^2 - a\zeta_2^2 - b(\zeta_3^2 - c\zeta_4^2)$$

with  $a, b, c \in F^\times$ . Evidently  $a$  and  $c$  are not squares, and if  $d$  is represented by  $\xi_3^2 - c\xi_4^2$ , then  $bd$  is not represented by  $\xi_1^2 - a\xi_2^2$ . Thus  $(c, d)_F = 1$  implies  $(a, bd)_F = -1$ . In particular,  $(a, b)_F = -1$  and hence  $(c, d)_F = 1$  implies  $(a, d)_F = 1$ . By interchanging the roles of  $\xi_1^2 - a\xi_2^2$  and  $\xi_3^2 - c\xi_4^2$ , we see that  $(a, d)_F = 1$  also implies  $(c, d)_F = 1$ . Hence  $(ac, d)_F = 1$  for all  $d \in F^\times$ . Thus  $ac$  is a square and  $g$  is equivalent to

$$f_{a,b} = \xi_1^2 - a\xi_2^2 - b(\xi_3^2 - a\xi_4^2).$$

We now show that  $f_{a,b}$  and  $f_{a',b'}$  are equivalent if  $(a, b)_F = (a', b')_F = -1$ . Suppose first that  $(a, b')_F = -1$ . Then  $(a, bb')_F = 1$  and there exist  $x_3, x_4 \in F$  such that  $b' = b(x_3^2 - ax_4^2)$ . Since

$$(x_3^2 - ax_4^2)(\xi_3^2 - a\xi_4^2) = \eta_3^2 - a\eta_4^2,$$

where  $\eta_3 = x_3\xi_3 + ax_4\xi_4$ ,  $\eta_4 = x_4\xi_3 + x_3\xi_4$ , it follows that  $f_{a,b'}$  is equivalent to  $f_{a,b}$ . For the same reason  $f_{a,b'}$  is equivalent to  $f_{a',b'}$  and thus  $f_{a,b}$  is equivalent to  $f_{a',b'}$ . By symmetry, the same conclusion holds if  $(a', b)_F = -1$ . Thus we now suppose

$$(a, b')_F = (a', b)_F = 1.$$

But then  $(a, bb')_F = (a', bb')_F = -1$  and so, by what we have already proved,

$$f_{a,b} \sim f_{a,bb'} \sim f_{a',bb'} \sim f_{a',b'}.$$

Together, the last two paragraphs show that if  $F$  is a Hilbert field, then all anisotropic quaternary quadratic forms which represent 1 are equivalent. Hence the Hilbert field  $F$  is of type (B) if every anisotropic quaternary quadratic form represents 1.

Suppose now that some anisotropic quaternary quadratic form does not represent 1. Then some scalar multiple of this form represents 1, but is not universal. Thus  $f_{a,b}$  is not universal for some  $a, b \in F^\times$  with  $(a, b)_F = -1$ . By Lemma 29, the set  $G_{a,b}$  of all  $c \in F^\times$  which are represented by  $f_{a,b}$  is a subgroup of  $F^\times$ . In fact  $G_{a,b} = G_a$ , since  $G_a \subseteq G_{a,b}$ ,  $G_{a,b} \neq F^\times$  and  $G_a$  has index 2 in  $F^\times$ . Since  $f_{a,b} \sim f_{b,a}$ , we have also  $G_{a,b} = G_b$ . Thus  $(a, c)_F = (b, c)_F$  for all  $c \in F^\times$ , and hence  $(ab, c)_F = 1$  for all  $c \in F^\times$ . Thus  $ab$  is a square and  $(a, a)_F = (a, b)_F = -1$ . Since  $(a, -a)_F = 1$ , it follows that  $(a, -1)_F = -1$ . Hence  $f_{a,b} \sim f_{a,a} \sim f_{a,-1}$ . Replacing  $a, b$  by  $-1, a$  we now obtain  $(-1, -1)_F = -1$  and  $f_{a,-1} \sim f_{-1,-1}$ .

Thus the form

$$f = \xi_1^2 + \xi_2^2 + \xi_3^2 + \xi_4^2$$

is not universal and the subgroup  $P$  of all elements of  $F^\times$  represented by  $f$  coincides with the set of all elements of  $F^\times$  represented by  $\xi^2 + \eta^2$ . Hence  $P + P \subseteq P$  and  $P$  is the set of all  $c \in F^\times$  such that  $(-1, c)_F = 1$ . Consequently  $-1 \notin P$  and  $F$  is the disjoint union of the sets  $\{0\}$ ,  $P$  and  $-P$ . Thus  $F$  is an ordered field with  $P$  as the set of positive elements.

For any  $c \in F^\times$ ,  $c^2 \in P$ . It follows that if  $a, b \in P$  then  $(-a, -b)_F = -1$ , since  $a\xi^2 + b\eta^2$  does not represent  $-1$ . Hence it follows that, if  $a, b \in P$ ,

then  $(-a, -b)_F = -1 = (-1, -b)_F$  and  $(-a, b)_F = 1 = (-1, b)_F$ . Thus, for all  $c \in F^\times$ ,  $(-a, c)_F = (-1, c)_F$  and hence  $(a, c)_F = 1$ . Therefore  $a$  is a square and the Hilbert field  $F$  is of type (A).  $\square$

**Proposition 31** *If  $F$  is a Hilbert field of type (B), then any quadratic form  $f$  in more than 4 variables is isotropic.*

*For any prime  $p$ , the field  $\mathbb{Q}_p$  of  $p$ -adic numbers is a Hilbert field of type (B).*

*Proof* The quadratic form  $f$  is equivalent to a diagonal form  $a_1\xi_1^2 + \cdots + a_n\xi_n^2$ , where  $n > 4$ . If  $g = a_1\xi_1^2 + \cdots + a_4\xi_4^2$  is isotropic, then so also is  $f$ . If  $g$  is anisotropic then, since  $F$  is of type (B), it is universal and represents  $-a_5$ . This proves the first part of the proposition.

We already know that  $\mathbb{Q}_p$  is a Hilbert field and we have already shown, after the proof of Corollary VI.17, that  $\mathbb{Q}_p$  is not an ordered field. Hence  $\mathbb{Q}_p$  is a Hilbert field of type (B).  $\square$

Proposition 10 shows that two non-singular quadratic forms in  $n$  variables, with coefficients from a Hilbert field of type (A), are equivalent over  $F$  if and only if they have the same positive index. We consider next the equivalence of quadratic forms with coefficients from a Hilbert field of type (B). We will show that they are classified by their determinant and their Hasse invariant.

If a non-singular quadratic form  $f$ , with coefficients from a Hilbert field  $F$ , is equivalent to a diagonal form  $a_1\xi_1^2 + \cdots + a_n\xi_n^2$ , then its *Hasse invariant* is defined to be the product of Hilbert symbols

$$s_F(f) = \prod_{1 \leq j < k \leq n} (a_j, a_k)_F.$$

We write  $s_p(f)$  for  $s_F(f)$  when  $F = \mathbb{Q}_p$ . (It should be noted that some authors define the Hasse invariant with  $\prod_{j \leq k}$  in place of  $\prod_{j < k}$ ). It must first be shown that this is indeed an invariant of  $f$ , and for this we make use of *Witt's chain equivalence theorem*:

**Lemma 32** *Let  $V$  be a non-singular quadratic space over an arbitrary field  $F$ . If  $\mathcal{B} = \{u_1, \dots, u_n\}$  and  $\mathcal{B}' = \{u'_1, \dots, u'_n\}$  are both orthogonal bases of  $V$ , then there exists a chain of orthogonal bases  $\mathcal{B}_0, \mathcal{B}_1, \dots, \mathcal{B}_m$ , with  $\mathcal{B}_0 = \mathcal{B}$  and  $\mathcal{B}_m = \mathcal{B}'$ , such that  $\mathcal{B}_{j-1}$  and  $\mathcal{B}_j$  differ by at most 2 vectors for each  $j \in \{1, \dots, m\}$ .*

*Proof* Since there is nothing to prove if  $\dim V = n \leq 2$ , we assume that  $n \geq 3$  and the result holds for all smaller values of  $n$ . Let  $p = p(\mathcal{B})$  be the number of nonzero coefficients in the representation of  $u'_1$  as a linear combination of  $u_1, \dots, u_n$ . Without loss of generality we may suppose

$$u'_1 = \sum_{j=1}^p a_j u_j,$$

where  $a_j \neq 0$  ( $1 \leq j \leq p$ ). If  $p = 1$ , we may replace  $u_1$  by  $u'_1$  and the result now follows by applying the induction hypothesis to the subspace of all vectors orthogonal to  $u'_1$ . Thus we now assume  $p \geq 2$ . We have

$$a_1^2(u_1, u_1) + \cdots + a_p^2(u_p, u_p) = (u'_1, u'_1) \neq 0,$$

and each summand on the left is nonzero. If the sum of the first two terms is zero, then  $p > 2$  and either the sum of the first and third terms is nonzero or the sum of the second and third terms is nonzero. Hence we may suppose without loss of generality that

$$a_1^2(u_1, u_1) + a_2^2(u_2, u_2) \neq 0.$$

If we put

$$v_1 = a_1 u_1 + a_2 u_2, \quad v_2 = u_1 + b u_2, \quad v_j = u_j \quad \text{for } 3 \leq j \leq n,$$

where  $b = -a_1(u_1, u_1)/a_2(u_2, u_2)$ , then  $\mathcal{B}_1 = \{v_1, \dots, v_n\}$  is an orthogonal basis and  $u'_1 = v_1 + a_3 v_3 + \cdots + a_p v_p$ . Thus  $p(\mathcal{B}_1) < p(\mathcal{B})$ . By replacing  $\mathcal{B}$  by  $\mathcal{B}_1$  and repeating the procedure, we must arrive after  $s < n$  steps at an orthogonal basis  $\mathcal{B}_s$  for which  $p(\mathcal{B}_s) = 1$ . The induction hypothesis can now be applied to  $\mathcal{B}_s$  in the same way as for  $\mathcal{B}$ .  $\square$

**Proposition 33** *Let  $F$  be a Hilbert field. If the non-singular diagonal forms  $a_1 \xi_1^2 + \cdots + a_n \xi_n^2$  and  $b_1 \xi_1^2 + \cdots + b_n \xi_n^2$  are equivalent over  $F$ , then*

$$\prod_{1 \leq j < k \leq n} (a_j, a_k)_F = \prod_{1 \leq j < k \leq n} (b_j, b_k)_F.$$

*Proof* Suppose first that  $n = 2$ . Since  $a_1 \xi_1^2 + a_2 \xi_2^2$  represents  $b_1$ ,  $\xi_1^2 + a_1^{-1} a_2 \xi_2^2$  represents  $a_1^{-1} b_1$  and hence  $(-a_1^{-1} a_2, a_1^{-1} b_1)_F = 1$ . Thus  $(a_1 b_1, -a_1 a_2 b_1)_F = 1$  and hence  $(a_1 b_1, a_2 b_1)_F = 1$ . But (Proposition 28 (ii)) the Hilbert symbol is multiplicative, since  $F$  is a Hilbert field. It follows that  $(a_1, a_2)_F (b_1, a_1 a_2 b_1)_F = 1$ . Since the determinants  $a_1 a_2$  and  $b_1 b_2$  are in the same square class, this implies  $(a_1, a_2)_F = (b_1, b_2)_F$ , as we wished to prove.

Suppose now that  $n > 2$ . Since the Hilbert symbol is symmetric, the product  $\prod_{1 \leq j < k \leq n} (a_j, a_k)_F$  is independent of the ordering of  $a_1, \dots, a_n$ . It follows from Lemma 32 that we may restrict attention to the case where  $a_1 \xi_1^2 + a_2 \xi_2^2$  is equivalent to  $b_1 \xi_1^2 + b_2 \xi_2^2$  and  $a_j = b_j$  for all  $j > 2$ . Then  $(a_1, a_2)_F = (b_1, b_2)_F$ , by what we have already proved, and it is enough to show that

$$(a_1, c)_F (a_2, c)_F = (b_1, c)_F (b_2, c)_F \quad \text{for any } c \in F^\times.$$

But this follows from the multiplicativity of the Hilbert symbol and the fact that  $a_1 a_2$  and  $b_1 b_2$  are in the same square class.  $\square$

Proposition 33 shows that the Hasse invariant is well-defined.

**Proposition 34** *Two non-singular quadratic forms in  $n$  variables, with coefficients from a Hilbert field  $F$  of type (B), are equivalent over  $F$  if and only if they have the same Hasse invariant and their determinants are in the same square class.*

*Proof* Only the sufficiency of the conditions needs to be proved. Since this is trivial for  $n = 1$ , we suppose first that  $n = 2$ . It is enough to show that if

$$f = a(\xi_1^2 + d\xi_2^2), \quad g = b(\eta_1^2 + d\eta_2^2),$$

where  $(a, ad)_F = (b, bd)_F$ , then  $f$  is equivalent to  $g$ . The hypothesis implies  $(-d, a)_F = (-d, b)_F$  and hence  $(-d, ab)_F = 1$ . Thus  $\xi_1^2 + d\xi_2^2$  represents  $ab$  and  $f$  represents  $b$ . Since  $\det f$  and  $\det g$  are in the same square class, it follows that  $f$  is equivalent to  $g$ .

Suppose next that  $n \geq 3$  and the result holds for all smaller values of  $n$ . Let  $f(\xi_1, \dots, \xi_n)$  and  $g(\eta_1, \dots, \eta_n)$  be non-singular quadratic forms with  $\det f = \det g = d$  and  $s_F(f) = s_F(g)$ . By Proposition 31, the quadratic form

$$h(\xi_1, \dots, \xi_n, \eta_1, \dots, \eta_n) = f(\xi_1, \dots, \xi_n) - g(\eta_1, \dots, \eta_n)$$

is isotropic and hence, by Proposition 7, there exists some  $a_1 \in F^\times$  which is represented by both  $f$  and  $g$ . Thus

$$f \sim a_1 \xi_1^2 + f^*, \quad g \sim a_1 \eta_1^2 + g^*,$$

where

$$f^* = a_2 \xi_2^2 + \dots + a_n \xi_n^2, \quad g^* = b_2 \eta_2^2 + \dots + b_n \eta_n^2.$$

Evidently  $\det f^*$  and  $\det g^*$  are in the same square class and  $s_F(f) = c s_F(f^*)$ ,  $s_F(g) = c' s_F(g^*)$ , where

$$c = (a_1, a_2 \cdots a_n)_F = (a_1, a_1)_F (a_1, d)_F = (a_1, b_2 \cdots b_n)_F = c'.$$

Hence  $s_F(f^*) = s_F(g^*)$ . It follows from the induction hypothesis that  $f^* \sim g^*$ , and so  $f \sim g$ .  $\square$

### 3 The Hasse–Minkowski Theorem

Let  $a, b, c$  be nonzero squarefree integers which are relatively prime in pairs. It was proved by Legendre (1785) that the equation

$$ax^2 + by^2 + cz^2 = 0$$

has a nontrivial solution in integers  $x, y, z$  if and only if  $a, b, c$  are not all of the same sign and the congruences

$$u^2 \equiv -bc \pmod{a}, \quad v^2 \equiv -ca \pmod{b}, \quad w^2 \equiv -ab \pmod{c}$$

are all soluble.

It was first completely proved by Gauss (1801) that every positive integer which is not of the form  $4^n(8k+7)$  can be represented as a sum of three squares. Legendre had given a proof, based on the assumption that if  $a$  and  $m$  are relatively prime positive integers, then the arithmetic progression

$$a, a+m, a+2m, \dots$$

contains infinitely many primes. Although his proof of this assumption was faulty, his intuition that it had a role to play in the arithmetic theory of quadratic forms

was inspired. The assumption was first proved by Dirichlet (1837) and will be referred to here as ‘Dirichlet’s theorem on primes in an arithmetic progression’. In the present chapter Dirichlet’s theorem will simply be assumed, but it will be proved (in a quantitative form) in Chapter X.

It was shown by Meyer (1884), although the published proof was incomplete, that a quadratic form in five or more variables with integer coefficients is isotropic if it is neither positive definite nor negative definite.

The preceding results are all special cases of the *Hasse–Minkowski theorem*, which is the subject of this section. Let  $\mathbb{Q}$  denote the field of rational numbers. By Ostrowski’s theorem (Proposition VI.4), the completions  $\mathbb{Q}_v$  of  $\mathbb{Q}$  with respect to an arbitrary absolute value  $|\cdot|_v$  are the field  $\mathbb{Q}_\infty = \mathbb{R}$  of real numbers and the fields  $\mathbb{Q}_p$  of  $p$ -adic numbers, where  $p$  is an arbitrary prime. The Hasse–Minkowski theorem has the following statement:

*A non-singular quadratic form  $f(\xi_1, \dots, \xi_n)$  with coefficients from  $\mathbb{Q}$  is isotropic in  $\mathbb{Q}$  if and only if it is isotropic in every completion of  $\mathbb{Q}$ .*

This concise statement contains, and to some extent conceals, a remarkable amount of information. (Its equivalence to Legendre’s theorem when  $n = 3$  may be established by elementary arguments.) The theorem was first stated and proved by Hasse (1923). Minkowski (1890) had derived necessary and sufficient conditions for the equivalence over  $\mathbb{Q}$  of two non-singular quadratic forms with rational coefficients by using known results on quadratic forms with integer coefficients. The role of  $p$ -adic numbers was taken by congruences modulo prime powers. Hasse drew attention to the simplifications obtained by studying from the outset quadratic forms over the field  $\mathbb{Q}$ , rather than the ring  $\mathbb{Z}$ , and soon afterwards (1924) he showed that the theorem continues to hold if the rational field  $\mathbb{Q}$  is replaced by an arbitrary algebraic number field (with its corresponding completions).

The condition in the statement of the theorem is obviously necessary and it is only its sufficiency which requires proof. Before embarking on this we establish one more property of the Hilbert symbol for the field  $\mathbb{Q}$  of rational numbers.

**Proposition 35** *For any  $a, b \in \mathbb{Q}^\times$ , the number of completions  $\mathbb{Q}_v$  for which one has  $(a, b)_v = -1$  (where  $v$  denotes either  $\infty$  or an arbitrary prime  $p$ ) is finite and even.*

*Proof* By Proposition 23, it is sufficient to establish the result when  $a$  and  $b$  are square-free integers such that  $ab$  is also square-free. Then  $(a, b)_r = 1$  for any odd prime  $r$  which does not divide  $ab$ , by Proposition 25. We wish to show that  $\prod_v (a, b)_v = 1$ . Since the Hilbert symbol is multiplicative, it is sufficient to establish this in the following special cases: for  $a = -1$  and  $b = -1, 2, p$ ; for  $a = 2$  and  $b = p$ ; for  $a = p$  and  $b = q$ , where  $p$  and  $q$  are distinct odd primes. But it follows from Propositions 24, 25 and 27 that

$$\prod_v (-1, -1)_v = (-1, -1)_\infty (-1, -1)_2 = (-1)(-1) = 1;$$

$$\prod_v (-1, 2)_v = (-1, 2)_\infty (-1, 2)_2 = 1 \cdot 1 = 1;$$

$$\prod_v (-1, p)_v = (-1, p)_p (-1, p)_2 = (-1/p)(-1)^{(p-1)/2},$$

$$\prod_v (2, p)_v = (2, p)_p (2, p)_2 = (2/p)(-1)^{(p^2-1)/8},$$

$$\prod_v (p, q)_v = (p, q)_p (p, q)_q (p, q)_2 = (q/p)(p/q)(-1)^{(p-1)(q-1)/4}.$$

Hence the proposition holds if and only if

$$(-1/p) = (-1)^{(p-1)/2}, (2/p) = (-1)^{(p^2-1)/8}, (q/p)(p/q) = (-1)^{(p-1)(q-1)/4}.$$

Thus it is actually equivalent to the law of quadratic reciprocity and its two ‘supplements’.  $\square$

We are now ready to prove the Hasse–Minkowski theorem:

**Theorem 36** *A non-singular quadratic form  $f(\xi_1, \dots, \xi_n)$  with rational coefficients is isotropic in  $\mathbb{Q}$  if and only if it is isotropic in every completion  $\mathbb{Q}_v$ .*

*Proof* We may assume that the quadratic form is diagonal:

$$f = a_1 \xi_1^2 + \dots + a_n \xi_n^2,$$

where  $a_k \in \mathbb{Q}^\times$  ( $k = 1, \dots, n$ ). Moreover, by replacing  $\xi_k$  by  $r_k \xi_k$ , we may assume that each coefficient  $a_k$  is a square-free integer.

The proof will be broken into three parts, according as  $n = 2$ ,  $n = 3$  or  $n \geq 4$ . The proofs for  $n = 2$  and  $n = 3$  are quite independent. The more difficult proof for  $n \geq 4$  uses induction on  $n$  and Dirichlet’s theorem on primes in an arithmetic progression.

(i)  $n = 2$ : We show first that if  $a \in \mathbb{Q}^\times$  is a square in  $\mathbb{Q}_v^\times$  for all  $v$ , then  $a$  is already a square in  $\mathbb{Q}^\times$ . Since  $a$  is a square in  $\mathbb{Q}_\infty^\times$ , we have  $a > 0$ . Let  $a = \prod_p p^{\alpha_p}$  be the factorization of  $a$  into powers of distinct primes, where  $\alpha_p \in \mathbb{Z}$  and  $\alpha_p \neq 0$  for at most finitely many primes  $p$ . Since  $|a|_p = p^{-\alpha_p}$  and  $a$  is a square in  $\mathbb{Q}_p$ ,  $\alpha_p$  must be even. But if  $\alpha_p = 2\beta_p$  then  $a = b^2$ , where  $b = \prod_p p^{\beta_p}$ .

Suppose now that  $f = a_1 \xi_1^2 + a_2 \xi_2^2$  is isotropic in  $\mathbb{Q}_v$  for all  $v$ . Then  $a := -a_1 a_2$  is a square in  $\mathbb{Q}_v$  for all  $v$  and hence, by what we have just proved,  $a$  is a square in  $\mathbb{Q}$ . But if  $a = b^2$ , then  $a_1 a_2^2 + a_2 b^2 = 0$  and thus  $f$  is isotropic in  $\mathbb{Q}$ .

(ii)  $n = 3$ : By replacing  $f$  by  $-a_3 f$  and  $\xi_3$  by  $a_3 \xi_3$ , we see that it is sufficient to prove the theorem for

$$f = a\xi^2 + b\eta^2 - \zeta^2,$$

where  $a$  and  $b$  are nonzero square-free integers. The quadratic form  $f$  is isotropic in  $\mathbb{Q}_v$  if and only if  $(a, b)_v = 1$ . If  $a = 1$  or  $b = 1$ , then  $f$  is certainly isotropic in  $\mathbb{Q}$ . Since  $f$  is not isotropic in  $\mathbb{Q}_\infty$  if  $a = b = -1$ , this proves the result if  $|ab| = 1$ . We will assume that the result does not hold for some pair  $a, b$  and derive a contradiction. Choose a pair  $a, b$  for which the result does not hold and for which  $|ab|$  has its minimum value. Then  $a \neq 1$ ,  $b \neq 1$  and  $|ab| \geq 2$ . Without loss of generality we may assume  $|a| \leq |b|$ , and then  $|b| \geq 2$ .

We are going to show that there exists an integer  $c$  such that  $c^2 \equiv a \pmod{b}$ . Since  $\pm b$  is a product of distinct primes, it is enough to show that the congruence  $x^2 \equiv a \pmod{p}$  is soluble for each prime  $p$  which divides  $b$  (by Corollary II.38). Since this is obvious if  $a \equiv 0$  or  $1 \pmod{p}$ , we may assume that  $p$  is odd and  $a$  not divisible by  $p$ . Then, since  $f$  is isotropic in  $\mathbb{Q}_p$ ,  $(a, b)_p = 1$ . Hence  $a$  is a square mod  $p$  by Proposition 25.

Consequently there exist integers  $c, d$  such that  $a = c^2 - bd$ . Moreover, by adding to  $c$  a suitable multiple of  $b$  we may assume that  $|c| \leq |b|/2$ . Then

$$|d| = |c^2 - a|/|b| \leq |b|/4 + 1 < |b|$$

and  $d \neq 0$ , since  $a$  is square-free and  $a \neq 1$ . We have

$$bd(a\xi^2 + b\eta^2 - \zeta^2) = aX^2 + dY^2 - Z^2,$$

where

$$X = c\xi + \zeta, \quad Y = b\eta, \quad Z = a\xi + c\zeta.$$

Moreover the linear transformation  $\xi, \eta, \zeta \rightarrow X, Y, Z$  is invertible in any field of zero characteristic, since  $c^2 - a \neq 0$ . Hence, since  $f$  is isotropic in  $\mathbb{Q}_v$  for all  $v$ , so also is  $g = a\xi^2 + d\eta^2 - \zeta^2$ . Since  $f$  is not isotropic in  $\mathbb{Q}$ , by hypothesis, neither is  $g$ . But this contradicts the original choice of  $f$ , since  $|ad| < |ab|$ .

It may be noted that for  $n = 3$  it need only be assumed that  $f$  is isotropic in  $\mathbb{Q}_p$  for all primes  $p$ . For the preceding proof used the fact that  $f$  is isotropic in  $\mathbb{Q}_\infty$  only to exclude from consideration the quadratic form  $-\xi^2 - \eta^2 - \zeta^2$  and this quadratic form is anisotropic also in  $\mathbb{Q}_2$ , by Proposition 27. In fact for  $n = 3$  it need only be assumed that  $f$  is isotropic in  $\mathbb{Q}_v$  for all  $v$  with at most one exception since, by Proposition 35, the number of exceptions must be even.

(iii)  $n \geq 4$ : We have

$$f = a_1\xi_1^2 + \cdots + a_n\xi_n^2,$$

where  $a_1, \dots, a_n$  are square-free integers. We write  $f = g - h$ , where

$$g = a_1\xi_1^2 + a_2\xi_2^2, \quad h = -a_3\xi_3^2 - \cdots - a_n\xi_n^2.$$

Let  $S$  be the finite set consisting of  $\infty$  and all primes  $p$  which divide  $2a_1 \cdots a_n$ . By Proposition 7, for each  $v \in S$  there exists  $c_v \in \mathbb{Q}_v^\times$  which is represented in  $\mathbb{Q}_v$  by both  $g$  and  $h$ . We will show that we can take  $c_v$  to be the same nonzero integer  $c$  for every  $v \in S$ .

Let  $v = p$  be a prime in  $S$ . By multiplying by a square in  $\mathbb{Q}_p^\times$  we may assume that  $c_p = p^{\varepsilon_p} c'_p$ , where  $\varepsilon_p = 0$  or  $1$  and  $|c'_p|_p = 1$ . If  $p$  is odd and if  $b_p$  is an integer such that  $|c_p - b_p|_p \leq p^{-\varepsilon_p - 1}$ , then  $|b_p|_p = |c_p|_p$  and  $|b_p c_p^{-1} - 1|_p \leq p^{-1}$ . Hence  $b_p c_p^{-1}$  is a square in  $\mathbb{Q}_p^\times$ , by Proposition VI.16, and we can replace  $c_p$  by  $b_p$ . Similarly if  $p = 2$  and if  $b_2$  is an integer such that  $|c_2 - b_2|_2 \leq 2^{-\varepsilon_2 - 3}$ , then  $|b_2|_2 = |c_2|_2$  and  $|b_2 c_2^{-1} - 1|_2 \leq 2^{-3}$ . Hence  $b_2 c_2^{-1}$  is a square in  $\mathbb{Q}_2^\times$  and we can replace  $c_2$  by  $b_2$ .

By the Chinese remainder theorem (Corollary II.38), the simultaneous congruences

$$c \equiv b_2 \pmod{2^{\varepsilon_2+3}}, c \equiv b_p \pmod{p^{\varepsilon_p+1}} \quad \text{for every odd } p \in S,$$

have a solution  $c \in \mathbb{Z}$ , that is uniquely determined mod  $m$ , where  $m = 4 \prod_{p \in S} p^{\varepsilon_p+1}$ . In exactly the same way as before we can replace  $b_p$  by  $c$  for all primes  $p \in S$ . By choosing  $c$  to have the same sign as  $c_\infty$ , we can take  $c_v = c$  for all  $v \in S$ .

If  $d = \prod_{p \in S} p^{\varepsilon_p}$  is the greatest common divisor of  $c$  and  $m$  then, by Dirichlet's theorem on primes in an arithmetic progression, there exists an integer  $k$  with the same sign as  $c$  such that

$$c/d + km/d = \pm q,$$

where  $q$  is a prime. If we put

$$a = c + km = \pm dq,$$

then  $q$  is the only prime divisor of  $a$  which is not in  $S$  and the quadratic forms

$$g^* = -a\zeta_0^2 + a_1\zeta_1^2 + a_2\zeta_2^2, \quad h^* = a_3\zeta_3^2 + \cdots + a_n\zeta_n^2 + a\zeta_{n+1}^2$$

are isotropic in  $\mathbb{Q}_v$  for every  $v \in S$ , since  $c^{-1}a$  is a square in  $\mathbb{Q}_v^\times$ .

For all primes  $p$  not in  $S$ , except  $p = q$ ,  $a$  is not divisible by  $p$ . Hence, by the definition of  $S$  and Corollary 26,  $g^*$  is isotropic in  $\mathbb{Q}_v$  for all  $v$ , except possibly  $v = q$ . Consequently, by the final remark of part (ii) of the proof,  $g^*$  is isotropic in  $\mathbb{Q}$ .

Suppose first that  $n = 4$ . In this case, in the same way,  $h^* = a_3\zeta_3^2 + a_4\zeta_4^2 + a\zeta_5^2$  is also isotropic in  $\mathbb{Q}$ . Hence, by Proposition 6, there exist  $y_1, \dots, y_4 \in \mathbb{Q}$  such that

$$a_1y_1^2 + a_2y_2^2 = a = -a_3y_3^2 - a_4y_4^2.$$

Thus  $f$  is isotropic in  $\mathbb{Q}$ .

Suppose next that  $n \geq 5$  and the result holds for all smaller values of  $n$ . Then the quadratic form  $h^*$  is isotropic in  $\mathbb{Q}_v$ , not only for  $v \in S$ , but for all  $v$ . For if  $p$  is a prime which is not in  $S$ , then  $a_3, a_4, a_5$  are not divisible by  $p$ . It follows from Corollary 26 that the quadratic form  $a_3\zeta_3^2 + a_4\zeta_4^2 + a_5\zeta_5^2$  is isotropic in  $\mathbb{Q}_p$ , and hence  $h^*$  is also. Since  $h^*$  is a non-singular quadratic form in  $n - 1$  variables, it follows from the induction hypothesis that  $h^*$  is isotropic in  $\mathbb{Q}$ . The proof can now be completed in the same way as for  $n = 4$ .  $\square$

**Corollary 37** *A non-singular rational quadratic form in  $n \geq 5$  variables is isotropic in  $\mathbb{Q}$  if and only if it is neither positive definite nor negative definite.*

*Proof* This follows at once from Theorem 36, on account of Propositions 10 and 31.  $\square$

**Corollary 38** *A non-singular quadratic form over the rational field  $\mathbb{Q}$  represents a nonzero rational number  $c$  in  $\mathbb{Q}$  if and only if it represents  $c$  in every completion  $\mathbb{Q}_v$ .*

*Proof* Only the sufficiency of the condition requires proof. But if the rational quadratic form  $f(\xi_1, \dots, \xi_n)$  represents  $c$  in  $\mathbb{Q}_v$  for all  $v$  then, by Theorem 36, the quadratic form

$$f^*(\xi_0, \xi_1, \dots, \xi_n) = -c\xi_0^2 + f(\xi_1, \dots, \xi_n)$$

is isotropic in  $\mathbb{Q}$ . Hence  $f$  represents  $c$  in  $\mathbb{Q}$ , by Proposition 6.  $\square$

**Proposition 39** *Two non-singular quadratic forms with rational coefficients are equivalent over  $\mathbb{Q}$  if and only if they are equivalent over all completions  $\mathbb{Q}_v$ .*

*Proof* Again only the sufficiency of the condition requires proof. Let  $f$  and  $g$  be non-singular rational quadratic forms in  $n$  variables which are equivalent over  $\mathbb{Q}_v$  for all  $v$ .

Suppose first that  $n = 1$  and that  $f = a\xi^2$ ,  $g = b\eta^2$ . By hypothesis, for every  $v$  there exists  $t_v \in \mathbb{Q}_v^\times$  such that  $b = at_v^2$ . Thus  $ba^{-1}$  is a square in  $\mathbb{Q}_v^\times$  for every  $v$ , and hence  $ba^{-1}$  is a square in  $\mathbb{Q}^\times$ , by part (i) of the proof of Theorem 36. Therefore  $f$  is equivalent to  $g$  over  $\mathbb{Q}$ .

Suppose now that  $n > 1$  and the result holds for all smaller values of  $n$ . Choose some  $c \in \mathbb{Q}^\times$  which is represented by  $f$  in  $\mathbb{Q}$ . Then  $f$  certainly represents  $c$  in  $\mathbb{Q}_v$  and hence  $g$  represents  $c$  in  $\mathbb{Q}_v$ , since  $g$  is equivalent to  $f$  over  $\mathbb{Q}_v$ . Since this holds for all  $v$ , it follows from Corollary 38 that  $g$  represents  $c$  in  $\mathbb{Q}$ .

Thus, by the remark after the proof of Proposition 2,  $f$  is equivalent over  $\mathbb{Q}$  to a quadratic form  $c\xi_1^2 + f^*(\xi_2, \dots, \xi_n)$  and  $g$  is equivalent over  $\mathbb{Q}$  to a quadratic form  $c\xi_1^2 + g^*(\xi_2, \dots, \xi_n)$ . Since  $f$  is equivalent to  $g$  over  $\mathbb{Q}_v$ , it follows from Witt's cancellation theorem that  $f^*(\xi_2, \dots, \xi_n)$  is equivalent to  $g^*(\xi_2, \dots, \xi_n)$  over  $\mathbb{Q}_v$ . Since this holds for every  $v$ , it follows from the induction hypothesis that  $f^*$  is equivalent to  $g^*$  over  $\mathbb{Q}$ , and so  $f$  is equivalent to  $g$  over  $\mathbb{Q}$ .  $\square$

**Corollary 40** *Two non-singular quadratic forms  $f$  and  $g$  in  $n$  variables with rational coefficients are equivalent over the rational field  $\mathbb{Q}$  if and only if*

- (i)  $(\det f)/(\det g)$  is a square in  $\mathbb{Q}^\times$ ,
- (ii)  $\text{ind}^+ f = \text{ind}^+ g$ ,
- (iii)  $s_p(f) = s_p(g)$  for every prime  $p$ .

*Proof* This follows at once from Proposition 39, on account of Propositions 10 and 34.  $\square$

The *strong Hasse principle* (Theorem 36) says that a quadratic form is *isotropic* over the global field  $\mathbb{Q}$  if (and only if) it is isotropic over all its local completions  $\mathbb{Q}_v$ . The so-named *weak Hasse principle* (Proposition 39) says that two quadratic forms are *equivalent* over  $\mathbb{Q}$  if (and only if) they are equivalent over all  $\mathbb{Q}_v$ . These *local-global principles* have proved remarkably fruitful. They organize the subject, they can be extended to other situations and, even when they fail, they are still a useful guide. We describe some results which illustrate these remarks.

As mentioned at the beginning of this section, the strong Hasse principle continues to hold when the rational field is replaced by any algebraic number field. Waterhouse (1976) has established the weak Hasse principle for pairs of quadratic forms: if over every completion  $\mathbb{Q}_v$  there is a change of variables taking both  $f_1$  to  $g_1$  and  $f_2$  to  $g_2$ , then there is also such a change of variables over  $\mathbb{Q}$ . For quadratic forms over the field

$F = K(t)$  of rational functions in one variable with coefficients from a field  $K$ , the weak Hasse principle always holds, and the strong Hasse principle holds for  $K = \mathbb{R}$ , but not for all fields  $K$ .

The strong Hasse principle also fails for polynomial forms over  $\mathbb{Q}$  of degree  $> 2$ . For example, Selmer (1951) has shown that the cubic equation  $3x^3 + 4y^3 + 5z^3 = 0$  has no nontrivial solutions in  $\mathbb{Q}$ , although it has nontrivial solutions in every completion  $\mathbb{Q}_v$ . However, Gusić (1995) has proved the weak Hasse principle for non-singular ternary cubic forms.

Finally, we draw attention to a remarkable local-global principle of Rumely (1986) for algebraic integer solutions of arbitrary systems of polynomial equations

$$f_1(\xi_1, \dots, \xi_n) = \dots = f_r(\xi_1, \dots, \xi_n) = 0$$

with rational coefficients.

We now give some applications of the results which have been established.

**Proposition 41** *A positive integer can be represented as the sum of the squares of three integers if and only if it is not of the form  $4^n b$ , where  $n \geq 0$  and  $b \equiv 7 \pmod{8}$ .*

*Proof* The necessity of the condition is easily established. Since the square of any integer is congruent to 0, 1 or 4 mod 8, the sum of three squares cannot be congruent to 7. For the same reason, if there exist integers  $x, y, z$  such that  $x^2 + y^2 + z^2 = 4^n b$ , where  $n \geq 1$  and  $b$  is odd, then  $x, y, z$  must all be even and thus  $(x/2)^2 + (y/2)^2 + (z/2)^2 = 4^{n-1}b$ . By repeating the argument  $n$  times, we see that there is no such representation if  $b \equiv 7 \pmod{8}$ .

We show next that any positive integer which satisfies this necessary condition is the sum of three squares of *rational* numbers. We need only show that any positive integer  $a \not\equiv 7 \pmod{8}$ , which is not divisible by 4, is represented in  $\mathbb{Q}$  by the quadratic form

$$f = \xi_1^2 + \xi_2^2 + \xi_3^2.$$

For every odd prime  $p$ ,  $f$  is isotropic in  $\mathbb{Q}_p$ , by Corollary 26, and hence any integer is represented in  $\mathbb{Q}_p$  by  $f$ , by Proposition 5. By Corollary 38, it only remains to show that  $f$  represents  $a$  in  $\mathbb{Q}_2$ .

It is easily seen that if  $a \equiv 1, 3$  or  $5 \pmod{8}$ , then there exist integers  $x_1, x_2, x_3 \in \{0, 1, 2\}$  such that

$$x_1^2 + x_2^2 + x_3^2 \equiv a \pmod{8}.$$

Hence  $a^{-1}(x_1^2 + x_2^2 + x_3^2)$  is a square in  $\mathbb{Q}_2^\times$  and  $f$  represents  $a$  in  $\mathbb{Q}_2$ .

Again, if  $a \equiv 2$  or  $6 \pmod{8}$ , then  $a \equiv 2, 6, 10$  or  $14 \pmod{2^4}$  and it is easily seen that there exist integers  $x_1, x_2, x_3 \in \{0, 1, 2, 3\}$  such that

$$x_1^2 + x_2^2 + x_3^2 \equiv a \pmod{2^4}.$$

Hence  $a^{-1}(x_1^2 + x_2^2 + x_3^2)$  is a square in  $\mathbb{Q}_2^\times$  and  $f$  represents  $a$  in  $\mathbb{Q}_2$ .

To complete the proof of the proposition we show, by an elegant argument due to Aubry (1912), that if  $f$  represents  $c$  in  $\mathbb{Q}$  then it also represents  $c$  in  $\mathbb{Z}$ .

Let

$$(x, y) = \{f(x + y) - f(x) - f(y)\}/2$$

be the symmetric bilinear form associated with  $f$ , so that  $f(x) = (x, x)$ , and assume there exists a point  $x \in \mathbb{Q}^3$  such that  $(x, x) = c \in \mathbb{Z}$ . If  $x \notin \mathbb{Z}^3$ , we can choose  $z \in \mathbb{Z}^3$  so that each coordinate of  $z$  differs in absolute value by at most  $1/2$  from the corresponding coordinate of  $x$ . Hence if we put  $y = x - z$ , then  $y \neq 0$  and  $0 < (y, y) \leq 3/4$ .

If  $x' = x - \lambda y$ , where  $\lambda = 2(x, y)/(y, y)$ , then  $x' \in \mathbb{Q}^3$  and  $(x', x') = (x, x) = c$ . Substituting  $y = x - z$ , we obtain

$$(y, y)x' = (y, y)x - 2(x, y)y = \{(z, z) - (x, x)\}x + 2\{(x, x) - (x, z)\}z.$$

If  $m > 0$  is the least common denominator of the coordinates of  $x$ , so that  $mx \in \mathbb{Z}^3$ , it follows that

$$m(y, y)x' = \{(z, z) - c\}mx + 2\{mc - (mx, z)\}z \in \mathbb{Z}^3.$$

But

$$m(y, y) = m\{(x, x) - 2(x, z) + (z, z)\} = mc - 2(mx, z) + m(z, z) \in \mathbb{Z}.$$

Thus if  $m' > 0$  is the least common denominator of the coordinates of  $x'$ , then  $m'$  divides  $m(y, y)$ . Hence  $m' \leq (3/4)m$ . If  $x' \notin \mathbb{Z}^3$ , we can repeat the argument with  $x$  replaced by  $x'$ . After performing the process finitely many times we must obtain a point  $x^* \in \mathbb{Z}^3$  such that  $(x^*, x^*) = c$ .  $\square$

As another application of the preceding results we now prove

**Proposition 42** *Let  $n, a, b$  be integers with  $n > 1$ . Then there exists a nonsingular  $n \times n$  rational matrix  $A$  such that*

$$A^t A = aI_n + bJ_n, \quad (3)$$

where  $J_n$  is the  $n \times n$  matrix with all entries 1, if and only if  $a > 0$ ,  $a + bn > 0$  and

(i) for  $n$  odd:  $a + bn$  is a square and the quadratic form

$$a\xi^2 + (-1)^{(n-1)/2}b\eta^2 - \zeta^2$$

is isotropic in  $\mathbb{Q}$ ;

(ii) for  $n$  even:  $a(a + bn)$  is a square and either  $n \equiv 0 \pmod{4}$ , or  $n \equiv 2 \pmod{4}$  and  $a$  is a sum of two squares.

*Proof* If we put

$$B = \begin{bmatrix} 1 & 1 & \dots & 1 & 1 \\ -1 & 1 & \dots & 1 & 1 \\ 0 & -2 & \dots & 1 & 1 \\ & \dots & \dots & \dots & \\ 0 & 0 & \dots & 1-n & 1 \end{bmatrix},$$

then  $D := B^t B$  and  $E := B^t J B$  are diagonal matrices:

$$D = \text{diag}[d_1, \dots, d_{n-1}, n], \quad E = \text{diag}[0, \dots, 0, n^2],$$

where  $d_j = j(j+1)$  for  $1 \leq j < n$ . Hence, if  $C = D^{-1} B^t A B$ , then

$$C^t D C = B^t A^t A B.$$

Thus the rational matrix  $A$  satisfies (3) if and only if the rational matrix  $C$  satisfies

$$C^t D C = aD + bE,$$

and consequently if and only if the diagonal quadratic forms

$$f = d_1 \zeta_1^2 + \dots + d_{n-1} \zeta_{n-1}^2 + n \zeta_n^2, \quad g = a d_1 \eta_1^2 + \dots + a d_{n-1} \eta_{n-1}^2 + n(a + bn) \eta_n^2$$

are equivalent over  $\mathbb{Q}$ .

We now apply Corollary 40. Since  $(\det g)/(\det f) = a^{n-1}(a + bn)$ , the condition that  $\det g / \det f$  be a square in  $\mathbb{Q}^\times$  means that  $a + bn$  is a nonzero square if  $n$  is odd and  $a(a + bn)$  is a nonzero square if  $n$  is even. Since  $\text{ind}^+ f = n$ , the condition that  $\text{ind}^+ g = \text{ind}^+ f$  means that  $a > 0$  and  $a + bn > 0$ . The relation  $s_p(g) = s_p(f)$  takes the form

$$\prod_{1 \leq i < j < n} (ad_i, ad_j)_p \prod_{1 \leq i < n} (ad_i, n(a + bn))_p = \prod_{1 \leq i < j < n} (d_i, d_j)_p \prod_{1 \leq i < n} (d_i, n)_p.$$

The multiplicativity and symmetry of the Hilbert symbol imply that

$$(ad_i, ad_j)_p = (a, a)_p (a, d_i d_j)_p (d_i, d_j)_p.$$

Since  $(a, a)_p = (a, -1)_p$ , it follows that  $s_p(g) = s_p(f)$  if and only if

$$(a, -1)_p^{(n-1)(n-2)/2} (a, n)_p^{n-1} \prod_{1 \leq i < n} (ad_i, a + bn)_p \prod_{1 \leq i < j < n} (a, d_i d_j)_p = 1.$$

But

$$\prod_{1 \leq i < j < n} d_i d_j = (d_1 \cdots d_{n-1})^{n-2}$$

and, by the definition of  $d_j$ ,  $d_1 \cdots d_{n-1}$  is in the same rational square class as  $n$ . Hence  $s_p(g) = s_p(f)$  if and only if

$$(a, -1)_p^{(n-1)(n-2)/2} (a, n)_p (an, a + bn)_p = 1. \quad (4)$$

If  $n$  is odd, then  $a + bn$  is a square and (4) reduces to  $(a, (-1)^{(n-1)/2} n)_p = 1$ . But, since  $a + bn$  is a square, the quadratic form  $a\zeta^2 + bn\eta^2 - \zeta^2$  is isotropic in  $\mathbb{Q}$  and thus  $(a, bn)_p = 1$  for all  $p$ . Hence  $(a, (-1)^{(n-1)/2} n)_p = 1$  for all  $p$  if and only if  $(a, (-1)^{(n-1)/2} b)_p = 1$  for all  $p$ . Since  $a > 0$ , this is equivalent to (i).

If  $n$  is even, then  $a(a + bn)$  is a square and (4) reduces to  $(a, (-1)^{(n-2)/2} a)_p = 1$ . Since  $a > 0$ , this holds for all  $p$  if and only if the ternary quadratic form

$$a\xi^2 + (-1)^{(n-2)/2}a\eta^2 - \zeta^2,$$

is isotropic in  $\mathbb{Q}$ . Thus it is certainly satisfied if  $n \equiv 0 \pmod{4}$ . If  $n \equiv 2 \pmod{4}$  it is satisfied if and only if the quadratic form  $\xi^2 + \eta^2 - a\zeta^2$  is isotropic. Thus it is satisfied if  $a$  is a sum of two squares. It is not satisfied if  $a$  is not a sum of two squares since then, by Proposition II.39, for some prime  $p \equiv 3 \pmod{4}$ , the highest power of  $p$  which divides  $a$  is odd and

$$(a, a)_p = (a, -1)_p = (p, -1)_p = (-1)^{(p-1)/2} = -1. \quad \square$$

It is worth noting that the last part of this proof shows that if a positive integer  $a$  is a sum of two rational squares, then it is also a sum of two squares of integers.

It follows at once from Proposition 42 that, for any positive integer  $n$ , there is an  $n \times n$  rational matrix  $A$  such that  $A^t A = nI_n$  if and only if either  $n$  is an odd square, or  $n \equiv 2 \pmod{4}$  and  $n$  is a sum of two squares, or  $n \equiv 0 \pmod{4}$  (the Hadamard matrix case).

In Chapter V we considered not only Hadamard matrices, but also designs. We now use Proposition 42 to derive the necessary conditions for the existence of square 2-designs which were obtained by Bruck, Ryser and Chowla (1949/50). Let  $v, k, \lambda$  be integers such that  $0 < \lambda < k < v$  and  $k(k-1) = \lambda(v-1)$ . Since  $k - \lambda + \lambda v = k^2$ , it follows from Proposition 42 that there exists a  $v \times v$  rational matrix  $A$  such that

$$A^t A = (k - \lambda)I_v + \lambda J_v$$

if and only if, either  $v$  is even and  $k - \lambda$  is a square, or  $v$  is odd and the quadratic form

$$(k - \lambda)\xi^2 + (-1)^{(v-1)/2}\lambda\eta^2 - \zeta^2$$

is isotropic in  $\mathbb{Q}$ .

A projective plane of order  $d$  corresponds to a  $(d^2 + d + 1, d + 1, 1)$  (square) 2-design. In this case Proposition 42 tells us that there is no projective plane of order  $d$  if  $d$  is not a sum of two squares and  $d \equiv 1$  or  $2 \pmod{4}$ . In particular, there is no projective plane of order 6.

The existence of projective planes of any prime power order follows from the existence of finite fields of any prime power order. (All known projective planes are of prime power order, but even for  $d = 9$  there are projective planes of the same order  $d$  which are not isomorphic.) Since there is no projective plane of order 6, the least order in doubt is  $d = 10$ . The condition derived from Proposition 42 is obviously satisfied in this case, since

$$10\xi^2 - \eta^2 - \zeta^2 = 0$$

has the solution  $\xi = \eta = 1, \zeta = 3$ . However, Lam, Thiel and Swiercz (1989) have announced that, nevertheless, there is no projective plane of order 10. The result was obtained by a search involving thousands of hours time on a supercomputer and does not appear to have been independently verified.

#### 4 Supplements

It was shown in the proof of Proposition 41 that if an integer can be represented as a sum of 3 squares of rational numbers, then it can be represented as a sum of 3 squares of integers. A similar argument was used by Cassels (1964) to show that if a polynomial can be represented as a sum of  $n$  squares of rational functions, then it can be represented as a sum of  $n$  squares of polynomials. This was immediately generalized by Pfister (1965) in the following way:

**Proposition 43** *For any field  $F$ , if there exist scalars  $\alpha_1, \dots, \alpha_n \in F$  and rational functions  $r_1(t), \dots, r_n(t) \in F(t)$  such that*

$$p(t) = \alpha_1 r_1(t)^2 + \dots + \alpha_n r_n(t)^2$$

*is a polynomial, then there exist polynomials  $p_1(t), \dots, p_n(t) \in F[t]$  such that*

$$p(t) = \alpha_1 p_1(t)^2 + \dots + \alpha_n p_n(t)^2.$$

*Proof* Suppose first that  $n = 1$ . We can write  $r_1(t) = p_1(t)/q_1(t)$ , where  $p_1(t)$  and  $q_1(t)$  are relatively prime polynomials and  $q_1(t)$  has leading coefficient 1. Since

$$p(t)q_1(t)^2 = \alpha_1 p_1(t)^2,$$

we must actually have  $q_1(t) = 1$ .

Suppose now that  $n > 1$  and the result holds for all smaller values of  $n$ . We may assume that  $\alpha_j \neq 0$  for all  $j$ , since otherwise the result follows from the induction hypothesis. Suppose first that the quadratic form

$$\phi = \alpha_1 \xi_1^2 + \dots + \alpha_n \xi_n^2$$

is isotropic over  $F$ . In this case there exists an invertible linear transformation  $\xi_j = \sum_{k=1}^n \tau_{jk} \eta_k$  with  $\tau_{jk} \in F$  ( $1 \leq j, k \leq n$ ) such that

$$\phi = \eta_1^2 - \eta_2^2 + \beta_3 \eta_3^2 + \dots + \beta_n \eta_n^2,$$

where  $\beta_j \in F$  for all  $j > 2$ . If we substitute

$$\eta_1 = \{p(t) + 1\}/2, \eta_2 = \{p(t) - 1\}/2, \eta_j = 0 \quad \text{for all } j > 2,$$

we obtain a representation for  $p(t)$  of the required form.

Thus we now suppose that  $\phi$  is anisotropic over  $F$ . This implies that  $\phi$  is also anisotropic over  $F(t)$ , since otherwise there would exist a nontrivial representation

$$\alpha_1 q_1(t)^2 + \dots + \alpha_n q_n(t)^2 = 0,$$

where  $q_j(t) \in F[t]$  ( $1 \leq j \leq n$ ), and by considering the terms of highest degree we would obtain a contradiction.

By hypothesis there exists a representation

$$p(t) = \alpha_1 \{f_1(t)/f_0(t)\}^2 + \dots + \alpha_n \{f_n(t)/f_0(t)\}^2,$$

where  $f_0(t), f_1(t), \dots, f_n(t) \in F[t]$ . Assume that  $f_0$  does not divide  $f_j$  for some  $j \in \{1, \dots, n\}$ . Then  $d := \deg f_0 > 0$  and we can write

$$f_j(t) = g_j(t)f_0(t) + h_j(t),$$

where  $g_j(t), h_j(t) \in F[t]$  and  $\deg h_j < d$  ( $1 \leq j \leq n$ ).

Let

$$(x, y) = \{\phi(x + y) - \phi(x) - \phi(y)\}/2$$

be the symmetric bilinear form associated with the quadratic form  $\phi$  and put

$$f = (f_1, \dots, f_n), \quad g = (g_1, \dots, g_n), \quad h = (h_1, \dots, h_n).$$

If

$$f_0^* = \{(g, g) - p\}f_0 - 2\{(f, g) - pf_0\}, \quad f^* = \{(g, g) - p\}f - 2\{(f, g) - pf_0\}g,$$

and  $f^* = (f_1^*, \dots, f_n^*)$ , then clearly  $f_0^*, f_1^*, \dots, f_n^* \in F[t]$ . Since  $(f, f) = pf_0^2$  and  $g = (f - h)/f_0$ , we can also write

$$f_0^* = (h, h)/f_0, \quad f^* = \{(h, h)f - 2(f, h)h\}/f_0^2.$$

It follows that  $\deg f_0^* < d$  and  $(f^*, f^*) = pf_0^{*2}$ . Also  $f_0^* \neq 0$ , since  $h \neq 0$  and  $\phi$  is anisotropic. Thus

$$p(t) = a_1\{f_1^*(t)/f_0^*(t)\}^2 + \dots + a_n\{f_n^*(t)/f_0^*(t)\}^2.$$

If  $f_0^*$  does not divide  $f_j^*$  for some  $j \in \{1, \dots, n\}$ , we can repeat the process. After at most  $d$  steps we must obtain a representation for  $p(t)$  of the required form.  $\square$

It was already known to Hilbert (1888) that there is no analogue of Proposition 43 for polynomials in more than one variable. Motzkin (1967) gave the simple example

$$p(x, y) = 1 - 3x^2y^2 + x^4y^2 + x^2y^4,$$

which is a sum of 4 squares in  $\mathbb{R}(x, y)$ , but is not a sum of any finite number of squares in  $\mathbb{R}[x, y]$ .

In the same paper in which he proved Proposition 43 Pfister introduced his *multiplicative forms*. The quadratic forms  $f_a, f_{a,b}$  in §2 are examples of such forms. Pfister (1966) used his multiplicative forms to obtain several new results on the structure of the Witt ring and then (1967) to give a strong solution to Hilbert's 17th Paris problem. We restrict attention here to the latter application.

Let  $g(x), h(x) \in \mathbb{R}[x]$  be polynomials in  $n$  variables  $x = (\xi_1, \dots, \xi_n)$  with real coefficients. The rational function  $f(x) = g(x)/h(x)$  is said to be *positive definite* if  $f(a) \geq 0$  for every  $a \in \mathbb{R}^n$  such that  $h(a) \neq 0$ . Hilbert's 17th problem asks if every positive definite rational function can be represented as a sum of squares:

$$f(x) = f_1(x)^2 + \dots + f_s(x)^2,$$

where  $f_1(x), \dots, f_s(x) \in \mathbb{R}(x)$ . The question was answered affirmatively by Artin (1927). Artin's solution allowed the number  $s$  of squares to depend on the function  $f$ , and left open the possibility that there might be no uniform bound. Pfister showed that one can always take  $s = 2^n$ .

Finally we mention a conjecture of Oppenheim (1929–1953), that if  $f(\xi_1, \dots, \xi_n)$  is a non-singular isotropic real quadratic form in  $n \geq 3$  variables, which is not a scalar multiple of a rational quadratic form, then  $f(\mathbb{Z}^n)$  is dense in  $\mathbb{R}$ , i.e. for each  $\alpha \in \mathbb{R}$  and  $\varepsilon > 0$  there exist  $z_1, \dots, z_n \in \mathbb{Z}$  such that  $|f(z_1, \dots, z_n) - \alpha| < \varepsilon$ . (It is not difficult to show that this is not always true for  $n = 2$ .) Raghunathan (1980) made a general conjecture about Lie groups, which he observed would imply Oppenheim's conjecture. Oppenheim's conjecture was then proved in this way by Margulis (1987), using deep results from the theory of Lie groups and ergodic theory. The full conjecture of Raghunathan has now also been proved by Ratner (1991).

## 5 Further Remarks

Lam [18] gives a good introduction to the arithmetic theory of quadratic spaces. The Hasse–Minkowski theorem is also proved in Serre [29]. Additional information is contained in the books of Cassels [4], Kitaoka [16], Milnor and Husemoller [20], O'Meara [22] and Scharlau [28].

Quadratic spaces were introduced (under the name 'metric spaces') by Witt [32]. This noteworthy paper also made several other contributions: Witt's cancellation theorem, the Witt ring, Witt's chain equivalence theorem and the Hasse invariant in its most general form (as described below). Quadratic spaces are treated not only in books on the arithmetic of quadratic forms, but also in works of a purely algebraic nature, such as Artin [1], Dieudonné [8] and Jacobson [15].

An important property of the Witt ring was established by Merkur'ev (1981). In one formulation it says that every element of order 2 in the *Brauer group* of a field  $F$  is represented by the Clifford algebra of some quadratic form over  $F$ . For a clear account, see Lewis [19].

Our discussion of Hilbert fields is based on Fröhlich [9]. It may be shown that any locally compact non-archimedean valued field is a Hilbert field. Fröhlich gives other examples, but rightly remarks that the notion of Hilbert field clarifies the structure of the theory, even if one is interested only in the  $p$ -adic case. (The name 'Hilbert field' is also given to fields for which Hilbert's irreducibility theorem is valid.)

In the study of quadratic forms over an arbitrary field  $F$ , the Hilbert symbol  $(a, b/F)$  is a generalized quaternion algebra (more strictly, an equivalence class of such algebras) and the Hasse invariant is a tensor product of Hilbert symbols. See, for example, Lam [18].

Hasse's original proof of the Hasse–Minkowski theorem is reproduced in Hasse [13]. In principle it is the same as that given here, using a reduction argument due to Lagrange for  $n = 3$  and Dirichlet's theorem on primes in an arithmetic progression for  $n \geq 4$ .

The book of Cassels contains a proof of Theorem 36 which does not use Dirichlet's theorem, but it uses intricate results on genera of quadratic forms and is

not so ‘clean’. However, Conway [6] has given an elementary approach to the *equivalence* of quadratic forms over  $\mathbb{Q}$  (Proposition 39 and Corollary 40).

The book of O’Meara gives a proof of the Hasse–Minkowski theorem over any algebraic number field which avoids Dirichlet’s theorem and is ‘cleaner’ than ours, but it uses deep results from *class field theory*. For the latter, see Cassels and Fröhlich [5], Garbanati [10] and Neukirch [21].

To determine if a rational quadratic form  $f(\xi_1, \dots, \xi_n) = \sum_{j,k=1}^n a_{jk} \xi_j \xi_k$  is isotropic by means of Theorem 36 one has to show that it is isotropic in infinitely many completions. Nevertheless, the problem is a finite one. Clearly one may assume that the coefficients  $a_{jk}$  are integers and, if the equation  $f(x_1, \dots, x_n) = 0$  has a non-trivial solution in rational numbers, then it also has a nontrivial solution in integers. But Cassels has shown by elementary arguments that if  $f(x_1, \dots, x_n) = 0$  for some  $x_j \in \mathbb{Z}$ , not all zero, then the  $x_j$  may be chosen so that

$$\max_{1 \leq j \leq n} |x_j| \leq (3H)^{(n-1)/2},$$

where  $H = \sum_{j,k=1}^n |a_{jk}|$ . See Lemma 8.1 in Chapter 6 of [4].

Williams [31] gives a sharper result for the ternary quadratic form

$$g(\zeta, \eta, \zeta) = a\zeta^2 + b\eta^2 + c\zeta^2,$$

where  $a, b, c$  are integers with greatest common divisor  $d > 0$ . If  $g(x, y, z) = 0$  for some integers  $x, y, z$ , not all zero, then these integers may be chosen so that

$$|x| \leq |bc|^{1/2}/d, |y| \leq |ca|^{1/2}/d, |z| \leq |ab|^{1/2}/d.$$

The necessity of the Bruck–Ryser–Chowla conditions for the existence of symmetric block designs may also be established in a more elementary way, without also proving their sufficiency for rational equivalence. See, for example, Beth *et al.* [2]. For the non-existence of a projective plane of order 10, see C. Lam [17].

For various manifestations of the local-global principle, see Waterhouse [30], Hsia [14], Gusić [12] and Green *et al.* [11].

The work of Pfister instigated a flood of papers on the algebraic theory of quadratic forms. The books of Lam and Scharlau give an account of these developments. For Hilbert’s 17th problem, see also Pfister [23], [24] and Rajwade [25].

Although a positive integer which is a sum of  $n$  rational squares is also a sum of  $n$  squares of integers, the same does not hold for higher powers. For example,

$$5906 = (149/17)^4 + (25/17)^4,$$

but there do not exist integers  $m, n$  such that  $5906 = m^4 + n^4$ , since  $9^4 > 5906$ ,  $2 \cdot 7^4 < 5906$  and  $5906 - 8^4 = 1810$  is not a fourth power. For the representation of a polynomial as a sum of squares of polynomials, see Rudin [27].

For Oppenheim’s conjecture, see Dani and Margulis [7], Borel [3] and Ratner [26].

## 6 Selected References

- [1] E. Artin, *Geometric algebra*, reprinted, Wiley, New York, 1988. [Original edition, 1957]

- [2] T. Beth, D. Jungnickel and H. Lenz, *Design theory*, 2nd ed., 2 vols., Cambridge University Press, 1999.
- [3] A. Borel, Values of indefinite quadratic forms at integral points and flows on spaces of lattices, *Bull. Amer. Math. Soc. (N.S.)* **32** (1995), 184–204.
- [4] J.W.S. Cassels, *Rational quadratic forms*, Academic Press, London, 1978.
- [5] J.W.S. Cassels and A. Fröhlich (ed.), *Algebraic number theory*, Academic Press, London, 1967.
- [6] J.H. Conway, Invariants for quadratic forms, *J. Number Theory* **5** (1973), 390–404.
- [7] S.G. Dani and G.A. Margulis, Values of quadratic forms at integral points: an elementary approach, *Enseign. Math.* **36** (1990), 143–174.
- [8] J. Dieudonné, *La géométrie des groupes classiques*, 2nd ed., Springer-Verlag, Berlin, 1963.
- [9] A. Fröhlich, Quadratic forms ‘à la’ local theory, *Proc. Camb. Phil. Soc.* **63** (1967), 579–586.
- [10] D. Garbanati, Class field theory summarized, *Rocky Mountain J. Math.* **11** (1981), 195–225.
- [11] B. Green, F. Pop and P. Roquette, On Rumely’s local-global principle, *Jahresber. Deutsch. Math.-Verein.* **97** (1995), 43–74.
- [12] I. Gusić, Weak Hasse principle for cubic forms, *Glas. Mat. Ser. III* **30** (1995), 17–24.
- [13] H. Hasse, *Mathematische Abhandlungen* (ed. H.W. Leopoldt and P. Roquette), Band I, de Gruyter, Berlin, 1975.
- [14] J.S. Hsia, On the Hasse principle for quadratic forms, *Proc. Amer. Math. Soc.* **39** (1973), 468–470.
- [15] N. Jacobson, *Basic Algebra I*, 2nd ed., Freeman, New York, 1985.
- [16] Y. Kitaoka, *Arithmetic of quadratic forms*, Cambridge University Press, 1993.
- [17] C.W.H. Lam, The search for a finite projective plane of order 10, *Amer. Math. Monthly* **98** (1991), 305–318.
- [18] T.Y. Lam, *The algebraic theory of quadratic forms*, revised 2nd printing, Benjamin, Reading, Mass., 1980.
- [19] D.W. Lewis, The Merkurjev–Suslin theorem, *Irish Math. Soc. Newsletter* **11** (1984), 29–37.
- [20] J. Milnor and D. Husemoller, *Symmetric bilinear forms*, Springer-Verlag, Berlin, 1973.
- [21] J. Neukirch, *Class field theory*, Springer-Verlag, Berlin, 1986.
- [22] O.T. O’Meara, *Introduction to quadratic forms*, corrected reprint, Springer-Verlag, New York, 1999. [Original edition, 1963]
- [23] A. Pfister, Hilbert’s seventeenth problem and related problems on definite forms, *Mathematical developments arising from Hilbert problems* (ed. F.E. Browder), pp. 483–489, Proc. Symp. Pure Math. **28**, Part 2, Amer. Math. Soc., Providence, Rhode Island, 1976.
- [24] A. Pfister, *Quadratic forms with applications to algebraic geometry and topology*, Cambridge University Press, 1995.
- [25] A.R. Rajwade, *Squares*, Cambridge University Press, 1993.
- [26] M. Ratner, Interactions between ergodic theory, Lie groups, and number theory, *Proceedings of the International Congress of Mathematicians: Zürich 1994*, pp. 157–182, Birkhäuser, Basel, 1995.
- [27] W. Rudin, Sums of squares of polynomials, *Amer. Math. Monthly* **107** (2000), 813–821.
- [28] W. Scharlau, *Quadratic and Hermitian forms*, Springer-Verlag, Berlin, 1985.
- [29] J.-P. Serre, *A course in arithmetic*, Springer-Verlag, New York, 1973.
- [30] W.C. Waterhouse, Pairs of quadratic forms, *Invent. Math.* **37** (1976), 157–164.
- [31] K.S. Williams, On the size of a solution of Legendre’s equation, *Utilitas Math.* **34** (1988), 65–72.
- [32] E. Witt, Theorie der quadratischen Formen in beliebigen Körpern, *J. Reine Angew. Math.* **176** (1937), 31–44.

## VIII

### The Geometry of Numbers

It was shown by Hermite (1850) that if

$$f(x) = x^t A x$$

is a positive definite quadratic form in  $n$  real variables, then there exists a vector  $x$  with integer coordinates, not all zero, such that

$$f(x) \leq c_n (\det A)^{1/n},$$

where  $c_n$  is a positive constant depending only on  $n$ . Minkowski (1891) found a new and more geometric proof of Hermite's result, which gave a much smaller value for the constant  $c_n$ . Soon afterwards (1893) he noticed that his proof was valid not only for an  $n$ -dimensional ellipsoid  $f(x) \leq \text{const.}$ , but for any convex body which was symmetric about the origin. This led him to a large body of results, to which he gave the somewhat paradoxical name 'geometry of numbers'. It seems fair to say that Minkowski was the first to realize the importance of convexity for mathematics, and it was in his lattice point theorem that he first encountered it.

#### 1 Minkowski's Lattice Point Theorem

A set  $C \subseteq \mathbb{R}^n$  is said to be *convex* if  $x_1, x_2 \in C$  implies  $\theta x_1 + (1 - \theta)x_2 \in C$  for  $0 < \theta < 1$ . Geometrically, this means that whenever two points belong to the set the whole line segment joining them is also contained in the set.

The *indicator function* or 'characteristic function' of a set  $S \subseteq \mathbb{R}^n$  is defined by  $\chi(x) = 1$  or  $0$  according as  $x \in S$  or  $x \notin S$ . If the indicator function is Lebesgue integrable, then the set  $S$  is said to have *volume*

$$\lambda(S) = \int_{\mathbb{R}^n} \chi(x) dx.$$

The indicator function of a convex set  $C$  is actually Riemann integrable. It is easily seen that if a convex set  $C$  is not contained in a hyperplane of  $\mathbb{R}^n$ , then its *interior* int  $C$  (see §4 of Chapter I) is not empty. It follows that  $\lambda(C) = 0$  if and only if  $C$  is

contained in a hyperplane, and  $0 < \lambda(C) < \infty$  if and only if  $C$  is bounded and is not contained in a hyperplane.

A set  $S \subseteq \mathbb{R}^n$  is said to be *symmetric* (with respect to the origin) if  $x \in S$  implies  $-x \in S$ . Evidently any (nonempty) symmetric convex set contains the origin.

A point  $x = (\xi_1, \dots, \xi_n) \in \mathbb{R}^n$  whose coordinates  $\xi_1, \dots, \xi_n$  are all integers will be called a *lattice point*. Thus the set of all lattice points in  $\mathbb{R}^n$  is  $\mathbb{Z}^n$ .

These definitions are the ingredients for Minkowski's *lattice point theorem*:

**Theorem 1** *Let  $C$  be a symmetric convex set in  $\mathbb{R}^n$ . If  $\lambda(C) > 2^n$ , or if  $C$  is compact and  $\lambda(C) = 2^n$ , then  $C$  contains a nonzero point of  $\mathbb{Z}^n$ .*

The proof of Theorem 1 will be deferred to §3. Here we illustrate the utility of the result by giving several applications, all of which go back to Minkowski himself.

**Proposition 2** *If  $A$  is an  $n \times n$  positive definite real symmetric matrix, then there exists a nonzero point  $x \in \mathbb{Z}^n$  such that*

$$x^t A x \leq c_n (\det A)^{1/n},$$

where  $c_n = (4/\pi) \{(n/2)!\}^{2/n}$ .

*Proof* For any  $\rho > 0$  the ellipsoid  $x^t A x \leq \rho$  is a compact symmetric convex set. By putting  $A = T^t T$ , for some nonsingular matrix  $T$ , it may be seen that the volume of this set is  $\kappa_n \rho^{n/2} (\det A)^{-1/2}$ , where  $\kappa_n$  is the volume of the  $n$ -dimensional unit ball. It follows from Theorem 1 that the ellipsoid contains a nonzero lattice point if  $\kappa_n \rho^{n/2} (\det A)^{-1/2} = 2^n$ . But, as we will see in §4 of Chapter IX,  $\kappa_n = \pi^{n/2}/(n/2)!$ , where  $x! = \Gamma(x+1)$ . This gives the value  $c_n$  for  $\rho$ .  $\square$

It follows from Stirling's formula (Chapter IX, §4) that  $c_n \sim 2n/\pi e$  for  $n \rightarrow \infty$ . Hermite had proved Proposition 2 with  $c_n = (4/3)^{(n-1)/2}$ . Hermite's value is smaller than Minkowski's for  $n \leq 8$ , but much larger for large  $n$ .

As a second application of Theorem 1 we prove Minkowski's *linear forms theorem*:

**Proposition 3** *Let  $A$  be an  $n \times n$  real matrix with determinant  $\pm 1$ . Then there exists a nonzero point  $x \in \mathbb{Z}^n$  such that  $Ax = y = (\eta_k)$  satisfies*

$$|\eta_1| \leq 1, \quad |\eta_k| < 1 \quad \text{for } 1 < k \leq n.$$

*Proof* For any positive integer  $m$ , let  $C_m$  be the set of all  $x \in \mathbb{R}^n$  such that  $Ax \in D_m$ , where

$$D_m = \{y = (\eta_k) \in \mathbb{R}^n : |\eta_1| \leq 1 + 1/m, |\eta_k| < 1 \text{ for } 2 \leq k \leq n\}.$$

Then  $C_m$  is a symmetric convex set, since  $A$  is linear and  $D_m$  is symmetric and convex. Moreover  $\lambda(C_m) = 2^n(1 + 1/m)$ , since  $\lambda(D_m) = 2^n(1 + 1/m)$  and  $A$  is volume-preserving. Therefore, by Theorem 1,  $C_m$  contains a lattice point  $x_m \neq O$ . Since  $C_m \subset C_1$  for all  $m > 1$  and the number of lattice points in  $C_1$  is finite, there exist only finitely many distinct points  $x_m$ . Thus there exists a lattice point  $x \neq O$  which belongs to  $C_m$  for infinitely many  $m$ . Evidently  $x$  has the required properties.  $\square$

The continued fraction algorithm enables one to find rational approximations to irrational numbers. The subject of *Diophantine approximation* is concerned with the more general problem of solving inequalities in integers. From Proposition 3 we can immediately obtain a result in this area due to Dirichlet (1842):

**Proposition 4** *Let  $A = (\alpha_{jk})$  be an  $n \times m$  real matrix and let  $t > 1$  be real. Then there exist integers  $q_1, \dots, q_m, p_1, \dots, p_n$ , with  $0 < \max(|q_1|, \dots, |q_m|) < t^{n/m}$ , such that*

$$\left| \sum_{k=1}^m \alpha_{jk} q_k - p_j \right| \leq 1/t \quad (1 \leq j \leq n).$$

*Proof* Since the matrix

$$\begin{pmatrix} t^{-n/m} I_m & 0 \\ tA & tI_n \end{pmatrix}$$

has determinant 1, it follows from Proposition 3 that there exists a nonzero vector

$$x = \begin{pmatrix} q \\ -p \end{pmatrix} \in \mathbb{Z}^{n+m}$$

such that

$$|q_k| < t^{n/m} \quad (k = 1, \dots, m),$$

$$\left| \sum_{k=1}^m \alpha_{jk} q_k - p_j \right| \leq 1/t \quad (j = 1, \dots, n).$$

Since  $q = 0$  would imply  $|p_j| < 1$  for all  $j$  and hence  $p = 0$ , which contradicts  $x \neq 0$ , we must have  $\max_k |q_k| > 0$ .  $\square$

**Corollary 5** *Let  $A = (\alpha_{jk})$  be an  $n \times m$  real matrix such that  $Az \notin \mathbb{Z}^n$  for any nonzero vector  $z \in \mathbb{Z}^m$ . Then there exist infinitely many  $(m+n)$ -tuples  $q_1, \dots, q_m, p_1, \dots, p_n$  of integers with greatest common divisor 1 and with arbitrarily large values of*

$$\|q\| = \max(|q_1|, \dots, |q_m|)$$

*such that*

$$\left| \sum_{k=1}^m \alpha_{jk} q_k - p_j \right| < \|q\|^{-m/n} \quad (1 \leq j \leq n).$$

*Proof* Let  $q_1, \dots, q_m, p_1, \dots, p_n$  be integers satisfying the conclusions of Proposition 4 for some  $t > 1$ . Evidently we may assume that  $q_1, \dots, q_m, p_1, \dots, p_n$  have no common divisor greater than 1. For given  $q_1, \dots, q_m$ , let  $\delta_j$  be the distance of  $\sum_{k=1}^m \alpha_{jk} q_k$  from the nearest integer and put  $\delta = \max \delta_j$  ( $1 \leq j \leq n$ ). By hypothesis  $0 < \delta < 1$ , and by construction

$$\delta \leq 1/t < \|q\|^{-m/n}.$$

Choosing some  $t' > 2/\delta$ , we find a new set of integers  $q'_1, \dots, q'_m, p'_1, \dots, p'_n$  satisfying the same requirements with  $t$  replaced by  $t'$ , and hence with  $\delta' \leq 1/t' < \delta/2$ . Proceeding in this way, we obtain a sequence of  $(m+n)$ -tuples of integers  $q_1^{(v)}, \dots, q_m^{(v)}, p_1^{(v)}, \dots, p_n^{(v)}$  for which  $\delta^{(v)} \rightarrow 0$  and hence  $\|q^{(v)}\| \rightarrow \infty$ , since we cannot have  $q^{(v)} = q$  for infinitely many  $v$ .  $\square$

The hypothesis of the corollary is certainly satisfied if  $1, \alpha_{j1}, \dots, \alpha_{jm}$  are linearly independent over the field  $\mathbb{Q}$  of rational numbers for some  $j \in \{1, \dots, n\}$ .

Minkowski also used his lattice point theorem to give the first proof that the discriminant of any algebraic number field, other than  $\mathbb{Q}$ , has absolute value greater than 1. The proof is given in most books on algebraic number theory.

## 2 Lattices

In the previous section we defined the set of lattice points to be  $\mathbb{Z}^n$ . However, this definition is tied to a particular coordinate system in  $\mathbb{R}^n$ . It is useful to consider lattices from a more intrinsic point of view. The key property is 'discreteness'.

With vector addition as the group operation,  $\mathbb{R}^n$  is an abelian group. A subgroup  $A$  is said to be *discrete* if there exists a ball with centre  $O$  which contains no other point of  $A$ . (More generally, a subgroup  $H$  of a topological group  $G$  is said to be discrete if there exists an open set  $U \subseteq G$  such that  $H \cap U = \{e\}$ , where  $e$  is the identity element of  $G$ .)

If  $A$  is a discrete subgroup of  $\mathbb{R}^n$ , then any bounded subset of  $\mathbb{R}^n$  contains at most finitely many points of  $A$  since, if there were infinitely many, they would have an accumulation point and their differences would accumulate at  $O$ . In particular,  $A$  is a closed subset of  $\mathbb{R}^n$ .

**Proposition 6** *If  $x_1, \dots, x_m$  are linearly independent vectors in  $\mathbb{R}^n$ , then the set*

$$A = \{\zeta_1 x_1 + \dots + \zeta_m x_m : \zeta_1, \dots, \zeta_m \in \mathbb{Z}\}$$

*is a discrete subgroup of  $\mathbb{R}^n$ .*

*Proof* It is clear that  $A$  is a subgroup of  $\mathbb{R}^n$ , since  $x, y \in A$  implies  $x - y \in A$ . If  $A$  is not discrete, then there exist  $y^{(v)} \in A$  with  $|y^{(1)}| > |y^{(2)}| > \dots$  and  $|y^{(v)}| \rightarrow 0$  as  $v \rightarrow \infty$ . Let  $V$  be the vector subspace of  $\mathbb{R}^n$  with basis  $x_1, \dots, x_m$  and for any vector

$$x = \alpha_1 x_1 + \dots + \alpha_m x_m,$$

where  $\alpha_k \in \mathbb{R}$  ( $1 \leq k \leq m$ ), put

$$\|x\| = \max(|\alpha_1|, \dots, |\alpha_m|).$$

This defines a norm on  $V$ . We have

$$y^{(v)} = \zeta_1^{(v)} x_1 + \dots + \zeta_m^{(v)} x_m,$$

where  $\zeta_k^{(v)} \in \mathbb{Z}$  ( $1 \leq k \leq m$ ). Since any two norms on a finite-dimensional vector space are equivalent (Lemma VI.7), it follows that  $\zeta_k^{(v)} \rightarrow 0$  as  $v \rightarrow \infty$  ( $1 \leq k \leq m$ ). Since  $\zeta_k^{(v)}$  is an integer, this is only possible if  $y^{(v)} = O$  for all large  $v$ , which is a contradiction.  $\square$

The converse of Proposition 6 is also valid. In fact we will prove a sharper result:

**Proposition 7** *If  $A$  is a discrete subgroup of  $\mathbb{R}^n$ , then there exist linearly independent vectors  $x_1, \dots, x_m$  in  $\mathbb{R}^n$  such that*

$$A = \{\zeta_1 x_1 + \dots + \zeta_m x_m : \zeta_1, \dots, \zeta_m \in \mathbb{Z}\}.$$

*Furthermore, if  $y_1, \dots, y_m$  is any maximal set of linearly independent vectors in  $A$ , we can choose  $x_1, \dots, x_m$  so that*

$$A \cap \langle y_1, \dots, y_k \rangle = \{\zeta_1 x_1 + \dots + \zeta_k x_k : \zeta_1, \dots, \zeta_k \in \mathbb{Z}\} \quad (1 \leq k \leq m),$$

*where  $\langle Y \rangle$  denotes the vector subspace generated by the set  $Y$ .*

*Proof* Let  $S_1$  denote the set of all  $\alpha_1 > 0$  such that  $\alpha_1 y_1 \in A$  and let  $\mu_1$  be the infimum of all  $\alpha_1 \in S_1$ . We are going to show that  $\mu_1 \in S_1$ . If this is not the case there exist  $\alpha_1^{(v)} \in S_1$  with  $\alpha_1^{(1)} > \alpha_1^{(2)} > \dots$  and  $\alpha_1^{(v)} \rightarrow \mu_1$  as  $v \rightarrow \infty$ . Since the ball  $|x| \leq (1 + \mu_1)|y_1|$  contains only finitely many points of  $A$ , this is a contradiction.

Any  $\alpha_1 \in S_1$  can be written in the form  $\alpha_1 = p\mu_1 + \theta$ , where  $p$  is a positive integer and  $0 \leq \theta < \mu_1$ . Since  $\theta > 0$  would imply  $\theta \in S_1$ , contrary to the definition of  $\mu_1$ , we must have  $\theta = 0$ . Hence if we put  $x_1 = \mu_1 y_1$ , then

$$A \cap \langle y_1 \rangle = \{\zeta_1 x_1 : \zeta_1 \in \mathbb{Z}\}.$$

Assume that, for some positive integer  $k$  ( $1 \leq k < m$ ), we have found vectors  $x_1, \dots, x_k \in A$  such that

$$A \cap \langle y_1, \dots, y_k \rangle = \{\zeta_1 x_1 + \dots + \zeta_k x_k : \zeta_1, \dots, \zeta_k \in \mathbb{Z}\}.$$

We will prove the proposition by showing that this assumption continues to hold when  $k$  is replaced by  $k + 1$ .

Any  $x \in A \cap \langle y_1, \dots, y_{k+1} \rangle$  has the form

$$x = \alpha_1 x_1 + \dots + \alpha_k x_k + \alpha_{k+1} y_{k+1},$$

where  $\alpha_1, \dots, \alpha_{k+1} \in \mathbb{R}$ . Let  $S_{k+1}$  denote the set of all  $\alpha_{k+1} > 0$  which arise in such representations and let  $\mu_{k+1}$  be the infimum of all  $\alpha_{k+1} \in S_{k+1}$ . We are going to show that  $\mu_{k+1} \in S_{k+1}$ . If  $\mu_{k+1} \notin S_{k+1}$ , there exist  $\alpha_{k+1}^{(v)} \in S_{k+1}$  with  $\alpha_{k+1}^{(1)} > \alpha_{k+1}^{(2)} > \dots$  and  $\alpha_{k+1}^{(v)} \rightarrow \mu_{k+1}$  as  $v \rightarrow \infty$ . Then  $A$  contains a point

$$x^{(v)} = \alpha_1^{(v)} x_1 + \dots + \alpha_k^{(v)} x_k + \alpha_{k+1}^{(v)} y_{k+1},$$

where  $\alpha_j^{(v)} \in \mathbb{R}$  ( $1 \leq j \leq k$ ). In fact, by subtracting an integral linear combination of  $x_1, \dots, x_k$  we may assume that  $0 \leq \alpha_j^{(v)} < 1$  ( $1 \leq j \leq k$ ). Since only finitely many points of  $A$  are contained in the ball  $|x| \leq |x_1| + \dots + |x_k| + (1 + \mu_{k+1})|y_{k+1}|$ , this is a contradiction.

Hence  $\mu_{k+1} > 0$  and  $A$  contains a vector

$$x_{k+1} = \alpha_1 x_1 + \dots + \alpha_k x_k + \mu_{k+1} y_{k+1}.$$

As for  $S_1$ , it may be seen that  $S_{k+1}$  consists of all positive integer multiples of  $\mu_{k+1}$ . Hence any  $x \in A \cap \langle y_1, \dots, y_{k+1} \rangle$  has the form

$$x = \zeta_1 x_1 + \dots + \zeta_k x_k + \zeta_{k+1} x_{k+1},$$

where  $\zeta_1, \dots, \zeta_k \in \mathbb{R}$  and  $\zeta_{k+1} \in \mathbb{Z}$ . Since

$$x - \zeta_{k+1} x_{k+1} \in A \cap \langle y_1, \dots, y_k \rangle,$$

we must actually have  $\zeta_1, \dots, \zeta_k \in \mathbb{Z}$ . □

By being more specific in the proof of Proposition 7 it may be shown that there is a *unique* choice of  $x_1, \dots, x_m$  such that

$$\begin{aligned} y_1 &= p_{11}x_1 \\ y_2 &= p_{21}x_1 + p_{22}x_2 \\ &\dots \\ y_m &= p_{m1}x_1 + p_{m2}x_2 + \dots + p_{mm}x_m, \end{aligned}$$

where  $p_{ij} \in \mathbb{Z}$ ,  $p_{ii} > 0$ , and  $0 \leq p_{ij} < p_{ii}$  if  $j < i$  (*Hermite's normal form*).

It is easily seen that in Proposition 7 we can choose  $x_i = y_i$  ( $1 \leq i \leq m$ ) if and only if, for any  $x \in A$  and any positive integer  $h$ ,  $x$  is an integral linear combination of  $y_1, \dots, y_m$  whenever  $hx$  is.

By combining Propositions 6 and 7 we obtain

**Proposition 8** *For a set  $A \subseteq \mathbb{R}^n$  the following two conditions are equivalent:*

- (i)  *$A$  is a discrete subgroup of  $\mathbb{R}^n$  and there exists  $R > 0$  such that, for each  $y \in \mathbb{R}^n$ , there is some  $x \in A$  with  $|y - x| < R$ ;*
- (ii) *there exist  $n$  linearly independent vectors  $x_1, \dots, x_n$  in  $\mathbb{R}^n$  such that*

$$A = \{\zeta_1 x_1 + \dots + \zeta_n x_n : \zeta_1, \dots, \zeta_n \in \mathbb{Z}\}.$$

*Proof* If (i) holds, then in the statement of Proposition 7 we must have  $m = n$ , i.e. (ii) holds. On the other hand, if (ii) holds then  $A$  is a discrete subgroup of  $\mathbb{R}^n$ , by Proposition 6. Moreover, for any  $y \in \mathbb{R}^n$  we can choose  $x \in A$  so that

$$y - x = \theta_1 x_1 + \dots + \theta_n x_n,$$

where  $0 \leq \theta_j < 1$  ( $j = 1, \dots, n$ ), and hence

$$|y - x| < |x_1| + \dots + |x_n|. \quad \square$$

A set  $A \subseteq \mathbb{R}^n$  satisfying either of the two equivalent conditions of Proposition 8 will be called a *lattice* and any element of  $A$  a *lattice point*. The vectors  $x_1, \dots, x_n$  in (ii) will be said to be a *basis* for the lattice.

A lattice is sometimes defined to be any discrete subgroup of  $\mathbb{R}^n$ , and what we have called a lattice is then called a 'nondegenerate' lattice. Our definition is chosen simply to avoid repetition of the word 'nondegenerate'. We may occasionally use the

more general definition and, with this warning, believe it will be clear from the context when this occurs.

The basis of a lattice is not uniquely determined. In fact  $y_1, \dots, y_n$  is also a basis if

$$y_j = \sum_{k=1}^n \alpha_{jk} x_k \quad (j = 1, \dots, n),$$

where  $A = (\alpha_{jk})$  is an  $n \times n$  matrix of integers such that  $\det A = \pm 1$ , since  $A^{-1}$  is then also a matrix of integers. Moreover, every basis  $y_1, \dots, y_n$  is obtained in this way. For if

$$y_j = \sum_{k=1}^n \alpha_{jk} x_k, \quad x_i = \sum_{j=1}^n \beta_{ij} y_j, \quad (i, j = 1, \dots, n),$$

where  $A = (\alpha_{jk})$  and  $B = (\beta_{ij})$  are  $n \times n$  matrices of integers, then  $BA = I$  and hence  $(\det B)(\det A) = 1$ . Since  $\det A$  and  $\det B$  are integers, it follows that  $\det A = \pm 1$ .

Let  $x_1, \dots, x_n$  be a basis for a lattice  $A \subseteq \mathbb{R}^n$ . If

$$x_k = \sum_{j=1}^n \gamma_{jk} e_j \quad (k = 1, \dots, n),$$

where  $e_1, \dots, e_n$  is the canonical basis for  $\mathbb{R}^n$  then, in terms of the nonsingular matrix  $T = (\gamma_{jk})$ , the lattice  $A$  is just the set of all vectors  $Tz$  with  $z \in \mathbb{Z}^n$ . The absolute value of the determinant of the matrix  $T$  does not depend on the choice of basis. For if  $x'_1, \dots, x'_n$  is any other basis, then

$$x'_i = \sum_{j=1}^n \alpha_{ij} x_j \quad (i = 1, \dots, n),$$

where  $A = (\alpha_{ij})$  is an  $n \times n$  matrix of integers with  $\det A = \pm 1$ . Thus

$$x'_k = \sum_{j=1}^n \gamma'_{jk} e_j \quad (k = 1, \dots, n),$$

where  $T' = (\gamma'_{jk})$  satisfies  $T' = TA^t$  and hence

$$|\det T'| = |\det T|.$$

The uniquely determined quantity  $|\det T|$  will be called the *determinant* of the lattice  $A$  and denoted by  $d(A)$ . (Some authors, e.g. Conway and Sloane [14], call  $|\det T|^2$  the determinant of  $A$ , but others prefer to call this the *discriminant* of  $A$ .)

The determinant  $d(A)$  has a simple geometrical interpretation. In fact it is the volume of the parallelootope  $\Pi$ , consisting of all points  $y \in \mathbb{R}^n$  such that

$$y = \theta_1 x_1 + \dots + \theta_n x_n,$$

where  $0 \leq \theta_k \leq 1$  ( $k = 1, \dots, n$ ). The interior of  $\Pi$  is a *fundamental domain* for the subgroup  $A$ , since

$$\mathbb{R}^n = \bigcup_{x \in A} (I + x),$$

$$\text{int}(I + x) \cap \text{int}(I + x') = \emptyset \quad \text{if } x, x' \in A \text{ and } x \neq x'.$$

For any lattice  $A \subseteq \mathbb{R}^n$ , the set  $A^*$  of all vectors  $y \in \mathbb{R}^n$  such that  $y^t x \in \mathbb{Z}$  for every  $x \in A$  is again a lattice, the *dual* (or ‘polar’ or ‘reciprocal’) of  $A$ . In fact,

$$\text{if } A = \{Tz : z \in \mathbb{Z}^n\}, \quad \text{then } A^* = \{(T^t)^{-1}w : w \in \mathbb{Z}^n\}.$$

Hence  $A$  is the dual of  $A^*$  and  $d(A)d(A^*) = 1$ . A lattice  $A$  is *self-dual* if  $A^* = A$ .

### 3 Proof of the Lattice Point Theorem; Other Results

In this section we take up the proof of Minkowski’s lattice point theorem. The proof will be based on a very general result, due to Blichfeldt (1914), which is not restricted to convex sets.

**Proposition 9** *Let  $S$  be a Lebesgue measurable subset of  $\mathbb{R}^n$ ,  $A$  a lattice in  $\mathbb{R}^n$  with determinant  $d(A)$  and  $m$  a positive integer.*

*If  $\lambda(S) > m d(A)$ , or if  $S$  is compact and  $\lambda(S) = m d(A)$ , then there exist  $m + 1$  distinct points  $x_1, \dots, x_{m+1}$  of  $S$  such that the differences  $x_j - x_k$  ( $1 \leq j, k \leq m + 1$ ) all lie in  $A$ .*

*Proof* Let  $b_1, \dots, b_n$  be a basis for  $A$  and let  $P$  be the half-open parallelotope consisting of all points  $x = \theta_1 b_1 + \dots + \theta_n b_n$ , where  $0 \leq \theta_i < 1$  ( $i = 1, \dots, n$ ). Then  $\lambda(P) = d(A)$  and

$$\mathbb{R}^n = \bigcup_{z \in A} (P + z), \quad (P + z) \cap (P + z') = \emptyset \quad \text{if } z \neq z'.$$

Suppose first that  $\lambda(S) > m d(A)$ . If we put

$$S_z = S \cap (P + z), \quad T_z = S_z - z,$$

then  $T_z \subseteq P$ ,  $\lambda(T_z) = \lambda(S_z)$  and

$$\lambda(S) = \sum_{z \in A} \lambda(S_z).$$

Hence

$$\sum_{z \in A} \lambda(T_z) = \lambda(S) > m d(A) = m \lambda(P).$$

Since  $T_z \subseteq P$  for every  $z$ , it follows that some point  $y \in P$  is contained in at least  $m + 1$  sets  $T_z$ . (In fact this must hold for all  $y$  in a subset of  $P$  of positive measure.) Thus there exist  $m + 1$  distinct points  $z_1, \dots, z_{m+1}$  of  $A$  and points  $x_1, \dots, x_{m+1}$  of  $S$  such that  $y = x_j - z_j$  ( $j = 1, \dots, m + 1$ ). Then  $x_1, \dots, x_{m+1}$  are distinct and

$$x_j - x_k = z_j - z_k \in A \quad (1 \leq j, k \leq m+1).$$

Suppose next that  $S$  is compact and  $\lambda(S) = m \, d(A)$ . Let  $\{\varepsilon_v\}$  be a decreasing sequence of positive numbers such that  $\varepsilon_v \rightarrow 0$  as  $v \rightarrow \infty$ , and let  $S_v$  denote the set of all points of  $\mathbb{R}^n$  distant at most  $\varepsilon_v$  from  $S$ . Then  $S_v$  is compact,  $\lambda(S_v) > \lambda(S)$  and

$$S_1 \supset S_2 \supset \cdots, \quad S = \bigcap_v S_v.$$

By what we have already proved, there exist  $m+1$  distinct points  $x_1^{(v)}, \dots, x_{m+1}^{(v)}$  of  $S_v$  such that  $x_j^{(v)} - x_k^{(v)} \in A$  for all  $j, k$ . Since  $S_v \subseteq S_1$  and  $S_1$  is compact, by restricting attention to a subsequence we may assume that  $x_j^{(v)} \rightarrow x_j$  as  $v \rightarrow \infty$  ( $j = 1, \dots, m+1$ ). Then  $x_j \in S$  and  $x_j^{(v)} - x_k^{(v)} \rightarrow x_j - x_k$ . Since  $x_j^{(v)} - x_k^{(v)} \in A$ , this is only possible if  $x_j - x_k = x_j^{(v)} - x_k^{(v)}$  for all large  $v$ . Hence  $x_1, \dots, x_{m+1}$  are distinct.  $\square$

Siegel (1935) has given an analytic formula which underlies Proposition 9 and enables it to be generalized. Although we will make no use of it, this formula will now be established. For notational simplicity we restrict attention to the (self-dual) lattice  $A = \mathbb{Z}^n$ .

**Proposition 10** *If  $\Psi : \mathbb{R}^n \rightarrow \mathbb{C}$  is a bounded measurable function which vanishes outside some compact set, then*

$$\int_{\mathbb{R}^n} \Psi(x) \overline{\phi(x)} dx = \sum_{w \in \mathbb{Z}^n} \left| \int_{\mathbb{R}^n} \Psi(x) e^{-2\pi i w^t x} dx \right|^2$$

where

$$\phi(x) = \sum_{z \in \mathbb{Z}^n} \Psi(x+z).$$

*Proof* Since  $\Psi$  vanishes outside a compact set, there exists a finite set  $T \subseteq \mathbb{Z}^n$  such that  $\Psi(x+z) = 0$  for all  $x \in \mathbb{R}^n$  if  $z \in \mathbb{Z}^n \setminus T$ . Thus the sum defining  $\phi(x)$  has only finitely many nonzero terms and  $\phi$  also is a bounded measurable function which vanishes outside some compact set.

If we write

$$x = (\zeta_1, \dots, \zeta_n), \quad z = (\zeta_1, \dots, \zeta_n),$$

then the sum defining  $\phi(x)$  is unaltered by the substitution  $\zeta_j \rightarrow \zeta_j + 1$  and hence  $\phi$  has period 1 in each of the variables  $\zeta_j$  ( $j = 1, \dots, n$ ). Let  $\Pi$  denote the fundamental parallelootope

$$\Pi = \{x = (\zeta_1, \dots, \zeta_n) \in \mathbb{R}^n : 0 \leq \zeta_j \leq 1 \text{ for } j = 1, \dots, n\}.$$

Since the functions  $e^{2\pi i w^t x}$  ( $w \in \mathbb{Z}^n$ ) are an orthogonal basis for  $L^2(\Pi)$ , Parseval's equality (Chapter I, §10) holds:

$$\int_{\Pi} |\phi(x)|^2 dx = \sum_{w \in \mathbb{Z}^n} |c_w|^2,$$

where

$$c_w = \int_{\Pi} \phi(x) e^{-2\pi i w^t x} dx.$$

But

$$\begin{aligned} c_w &= \int_{\Pi} \sum_{z \in \mathbb{Z}^n} \Psi(x+z) e^{-2\pi i w^t x} dx \\ &= \int_{\Pi} \sum_{z \in \mathbb{Z}^n} \Psi(x+z) e^{-2\pi i w^t (x+z)} dx, \end{aligned}$$

since  $e^{2k\pi i} = 1$  for any integer  $k$ . Hence

$$c_w = \int_{\mathbb{R}^n} \Psi(y) e^{-2\pi i w^t y} dy.$$

On the other hand,

$$\begin{aligned} \int_{\Pi} |\phi(x)|^2 dx &= \int_{\Pi} \sum_{z', z'' \in \mathbb{Z}^n} \Psi(x+z') \overline{\Psi(x+z'')} dx \\ &= \int_{\Pi} \sum_{z, z' \in \mathbb{Z}^n} \Psi(x+z') \overline{\Psi(x+z'+z)} dx \\ &= \int_{\mathbb{R}^n} \sum_{z \in \mathbb{Z}^n} \Psi(y) \overline{\Psi(y+z)} dy = \int_{\mathbb{R}^n} \Psi(y) \overline{\phi(y)} dy. \end{aligned}$$

Substituting these expressions in Parseval's equality, we obtain the result.  $\square$

Suppose, in particular, that  $\Psi$  takes only real nonnegative values. Then so also does  $\phi$  and

$$\int_{\mathbb{R}^n} \Psi(x) \phi(x) dx \leq \sup_{x \in \mathbb{R}^n} \phi(x) \int_{\mathbb{R}^n} \Psi(x) dx.$$

On the other hand, omitting all terms with  $w \neq 0$  we obtain

$$\sum_{w \in \mathbb{Z}^n} \left| \int_{\mathbb{R}^n} \Psi(x) e^{-2\pi i w^t x} dx \right|^2 \geq \left( \int_{\mathbb{R}^n} \Psi(x) dx \right)^2.$$

Hence, by Proposition 10,

$$\sup_{x \in \mathbb{R}^n} \phi(x) \geq \int_{\mathbb{R}^n} \Psi(x) dx.$$

For example, let  $S \subseteq \mathbb{R}^n$  be a measurable set with  $\lambda(S) > m$ . Then there exists a *bounded* measurable set  $S' \subseteq S$  with  $\lambda(S') > m$ . If we take  $\Psi$  to be the indicator function of  $S'$ , then

$$\int_{\mathbb{R}^n} \Psi(x) dx = \lambda(S') > m$$

and we conclude that there exists  $y \in \mathbb{R}^n$  such that

$$\sum_{z \in \mathbb{Z}^n} \Psi(y + z) = \phi(y) > m.$$

Since the only possible values of the summands on the left are 0 and 1, it follows that there exist  $m + 1$  distinct points  $z_1, \dots, z_{m+1} \in \mathbb{Z}^n = \Lambda$  such that  $y + z_j \in S$  for all  $j$ . The proof of Proposition 9 can now be completed in the same way as before.

Let  $\{K_\alpha\}$  be a family of subsets of  $\mathbb{R}^n$ , where each  $K_\alpha$  is the *closure* of a nonempty open set  $G_\alpha$ , i.e.  $K_\alpha$  is the intersection of all closed sets containing  $G_\alpha$ . The family  $\{K_\alpha\}$  is said to be a *packing* of  $\mathbb{R}^n$  if  $\alpha \neq \alpha'$  implies  $G_\alpha \cap G_{\alpha'} = \emptyset$  and is said to be a *covering* of  $\mathbb{R}^n$  if  $\mathbb{R}^n = \bigcup_\alpha K_\alpha$ . It is said to be a *tiling* of  $\mathbb{R}^n$  if it is both a packing and a covering.

For example, if  $\Pi$  is a fundamental parallelootope of a lattice  $\Lambda$ , then the family  $\{\Pi + a : a \in \Lambda\}$  is a tiling of  $\mathbb{R}^n$ . More generally, if  $G$  is a nonempty open subset of  $\mathbb{R}^n$  with closure  $K$ , we may ask whether the family  $\{K + a : a \in \Lambda\}$  of all  $\Lambda$ -translates of  $K$  is either a packing or a covering of  $\mathbb{R}^n$ . Some necessary conditions may be derived with the aid of Proposition 9:

**Proposition 11** *Let  $K$  be the closure of a bounded nonempty open set  $G \subseteq \mathbb{R}^n$  and let  $\Lambda$  be a lattice in  $\mathbb{R}^n$ .*

*If the  $\Lambda$ -translates of  $K$  are a covering of  $\mathbb{R}^n$  then  $\lambda(K) \geq d(\Lambda)$ , and the inequality is strict if they are not also a packing.*

*If the  $\Lambda$ -translates of  $K$  are a packing of  $\mathbb{R}^n$  then  $\lambda(K) \leq d(\Lambda)$ , and the inequality is strict if they are not also a covering.*

*Proof* Suppose first that the  $\Lambda$ -translates of  $K$  cover  $\mathbb{R}^n$ . Then every point of a fundamental parallelootope  $\Pi$  of  $\Lambda$  has the form  $x - a$ , where  $x \in K$  and  $a \in \Lambda$ . Hence

$$\begin{aligned} \lambda(K) &= \sum_{a \in \Lambda} \lambda(K \cap (\Pi + a)) \\ &= \sum_{a \in \Lambda} \lambda((K - a) \cap \Pi) \geq \lambda(\Pi) = d(\Lambda). \end{aligned}$$

Suppose, in addition, that the  $\Lambda$ -translates of  $K$  are not a packing of  $\mathbb{R}^n$ . Then there exist distinct points  $x_1, x_2$  in the interior  $G$  of  $K$  such that  $a = x_1 - x_2 \in \Lambda$ . Let

$$B_\varepsilon = \{x \in \mathbb{R}^n : |x| \leq \varepsilon\}.$$

We can choose  $\varepsilon > 0$  so small that the balls  $B_\varepsilon + x_1$  and  $B_\varepsilon + x_2$  are disjoint and contained in  $G$ . Then  $G' = G \setminus (B_\varepsilon + x_1)$  is a bounded nonempty open set with closure  $K' = K \setminus (\text{int} B_\varepsilon + x_1)$ . Since

$$B_\varepsilon + x_1 = B_\varepsilon + x_2 + a \subseteq K' + a,$$

the  $A$ -translates of  $K'$  contain  $K$  and therefore also cover  $\mathbb{R}^n$ . Hence, by what we have already proved,  $\lambda(K') \geq d(A)$ . Since  $\lambda(K) > \lambda(K')$ , it follows that  $\lambda(K) > d(A)$ .

Suppose now that the  $A$ -translates of  $K$  are a packing of  $\mathbb{R}^n$ . Then  $A$  does not contain the difference of two distinct points in the interior  $G$  of  $K$ , since  $G + a$  and  $G + b$  are disjoint if  $a, b$  are distinct points of  $A$ . It follows from Proposition 9 that

$$\lambda(K) = \lambda(G) \leq d(A).$$

Suppose, in addition, that the  $A$ -translates of  $K$  do not cover  $\mathbb{R}^n$ . Thus there exists a point  $y \in \mathbb{R}^n$  which is not in any  $A$ -translate of  $K$ . We will show that we can choose  $\varepsilon > 0$  so small that  $y$  is not in any  $A$ -translate of  $K + B_\varepsilon$ .

If this is not the case then, for any positive integer  $v$ , there exists  $a_v \in A$  such that

$$y \in K + B_{1/v} + a_v.$$

Evidently the sequence  $a_v$  is bounded and hence there exists  $a \in A$  such that  $a_v = a$  for infinitely many  $v$ . But then  $y \in K + a$ , which is contrary to hypothesis.

We may in addition assume  $\varepsilon$  chosen so small that  $|x| > 2\varepsilon$  for every nonzero  $x \in A$ . Then the set  $S = G \cup (B_\varepsilon + y)$  has the property that  $A$  does not contain the difference of any two distinct points of  $S$ . Hence, by Proposition 9,  $\lambda(S) \leq d(A)$ . Since

$$\lambda(K) = \lambda(G) < \lambda(S),$$

it follows that  $\lambda(K) < d(A)$ . □

We next apply Proposition 9 to convex sets. Minkowski's lattice point theorem (Theorem 1) is the special case  $m = 1$  (and  $A = \mathbb{Z}^n$ ) of the following generalization, due to van der Corput (1936):

**Proposition 12** *Let  $C$  be a symmetric convex subset of  $\mathbb{R}^n$ ,  $A$  a lattice in  $\mathbb{R}^n$  with determinant  $d(A)$ , and  $m$  a positive integer.*

*If  $\lambda(C) > 2^m m d(A)$ , or if  $C$  is compact and  $\lambda(C) = 2^m m d(A)$ , then there exist  $2m$  distinct nonzero points  $\pm y_1, \dots, \pm y_m$  of  $A$  such that*

$$\begin{aligned} y_j &\in C \quad (1 \leq j \leq m), \\ y_j - y_k &\in C \quad (1 \leq j, k \leq m). \end{aligned}$$

*Proof* The set  $S = \{x/2 : x \in C\}$  has measure  $\lambda(S) = \lambda(C)/2^n$ . Hence, by Proposition 9, there exist  $m + 1$  distinct points  $x_1, \dots, x_{m+1} \in C$  such that  $(x_j - x_k)/2 \in A$  for all  $j, k$ .

The vectors of  $\mathbb{R}^n$  may be totally ordered by writing  $x > x'$  if  $x - x'$  has its first nonzero coordinate positive. We assume the points  $x_1, \dots, x_{m+1} \in C$  numbered so that

$$x_1 > x_2 > \dots > x_{m+1}.$$

Put

$$y_j = (x_j - x_{m+1})/2 \quad (j = 1, \dots, m).$$

Then, by construction,  $y_j \in A$  ( $j = 1, \dots, m$ ). Moreover  $y_j \in C$ , since  $x_1, \dots, x_{m+1} \in C$  and  $C$  is symmetric, and similarly  $y_j - y_k = (x_j - x_k)/2 \in C$ . Finally, since

$$y_1 > y_2 > \dots > y_m > O,$$

we have  $y_j \neq O$  and  $y_j \neq \pm y_k$  if  $j \neq k$ . □

The conclusion of Proposition 12 need no longer hold if  $C$  is not compact and  $\lambda(C) = 2^n m d(A)$ . For example, take  $A = \mathbb{Z}^n$  and let  $C$  be the symmetric convex set

$$C = \{x = (\xi_1, \dots, \xi_n) \in \mathbb{R}^n : |\xi_1| < m, |\xi_j| < 1 \text{ for } 2 \leq j \leq n\}.$$

Then  $d(A) = 1$  and  $\lambda(C) = 2^n m$ . However, the only nonzero points of  $A$  in  $C$  are the  $2(m-1)$  points  $(\pm k, 0, \dots, 0)$  ( $1 \leq k \leq m-1$ ).

To provide a broader view of the geometry of numbers we now mention without proof some further results. A different generalization of Minkowski's lattice point theorem was already proved by Minkowski himself. Let  $A$  be a lattice in  $\mathbb{R}^n$  and let  $K$  be a compact symmetric convex subset of  $\mathbb{R}^n$  with nonempty interior. Then  $\rho K$  contains no nonzero point of  $A$  for small  $\rho > 0$  and contains  $n$  linearly independent points of  $A$  for large  $\rho > 0$ . Let  $\mu_i$  denote the infimum of all  $\rho > 0$  such that  $\rho K$  contains at least  $i$  linearly independent points of  $A$  ( $i = 1, \dots, n$ ). Clearly the *successive minima*  $\mu_i = \mu_i(K, A)$  satisfy the inequalities

$$0 < \mu_1 \leq \mu_2 \leq \dots \leq \mu_n < \infty.$$

Minkowski's lattice point theorem says that

$$\mu_1^n \lambda(K) \leq 2^n d(A).$$

Minkowski's *theorem on successive minima* strengthens this to

$$2^n d(A)/n! \leq \mu_1 \mu_2 \dots \mu_n \lambda(K) \leq 2^n d(A).$$

The lower bound is quite easy to prove, but the upper bound is more deep-lying — notwithstanding simplifications of Minkowski's original proof. If  $A = \mathbb{Z}^n$ , then equality holds in the upper bound for the *cube*  $K = \{(\xi_1, \dots, \xi_n) \in \mathbb{R}^n : |\xi_i| \leq 1 \text{ for all } i\}$  and in the lower bound for the *cross-polytope*  $K = \{(\xi_1, \dots, \xi_n) \in \mathbb{R}^n : \sum_{i=1}^n |\xi_i| \leq 1\}$ .

If  $K$  is a compact symmetric convex subset of  $\mathbb{R}^n$  with nonempty interior, we define its *critical determinant*  $\Delta(K)$  to be the infimum, over all lattices  $A$  with no nonzero point in the interior of  $K$ , of their determinants  $d(A)$ . A lattice  $A$  for which  $d(A) = \Delta(K)$  is called a *critical lattice* for  $K$ . It will be shown in §6 that a critical lattice always exists.

It follows from Proposition 12 that  $\Delta(K) \geq 2^{-n}\lambda(K)$ . A conjectured sharpening of Minkowski's theorem on successive minima, which has been proved by Minkowski (1896) himself for  $n = 2$  and for  $n$ -dimensional ellipsoids, and by Woods (1956) for  $n = 3$ , claims that

$$\mu_1\mu_2\cdots\mu_n\Delta(K) \leq d(\Delta).$$

The successive minima of a convex body are connected with those of its dual body. If  $K$  is a compact symmetric convex subset of  $\mathbb{R}^n$  with nonempty interior, then its *dual*

$$K^* = \{y \in \mathbb{R}^n : y^t x \leq 1 \text{ for all } x \in K\}$$

has the same properties, and  $K$  is the dual of  $K^*$ . Mahler (1939) showed that the successive minima of the dual body  $K^*$  with respect to the dual lattice  $\Delta^*$  are related to the successive minima of  $K$  with respect to  $\Delta$  by the inequalities

$$1 \leq \mu_i(K, \Delta)\mu_{n-i+1}(K^*, \Delta^*) \quad (i = 1, \dots, n),$$

and hence, by applying Minkowski's theorem on successive minima also to  $K^*$  and  $\Delta^*$ , he obtained inequalities in the opposite direction:

$$\mu_i(K, \Delta)\mu_{n-i+1}(K^*, \Delta^*) \leq 4^n/\lambda(K)\lambda(K^*) \quad (i = 1, \dots, n).$$

By further proving that  $\lambda(K)\lambda(K^*) \geq 4^n(n!)^{-2}$ , he deduced that

$$\mu_i(K, \Delta)\mu_{n-i+1}(K^*, \Delta^*) \leq (n!)^2 \quad (i = 1, \dots, n).$$

Dramatic improvements of these bounds have recently been obtained. Banaszczyk (1996), with the aid of techniques from harmonic analysis, has shown that there is a numerical constant  $C > 0$  such that, for all  $n \geq 1$  and all  $i \in \{1, \dots, n\}$ ,

$$\mu_i(K, \Delta)\mu_{n-i+1}(K^*, \Delta^*) \leq Cn(1 + \log n).$$

He had shown already (1993) that if  $K = B_1$  is the  $n$ -dimensional closed unit ball, which is self-dual, then for all  $n \geq 1$  and all  $i \in \{1, \dots, n\}$ ,

$$\mu_i(B_1, \Delta)\mu_{n-i+1}(B_1, \Delta^*) \leq n.$$

This result is close to being best possible, since there exists a numerical constant  $C' > 0$  and self-dual lattices  $\Delta_n \subseteq \mathbb{R}^n$  such that

$$\mu_1(B_1, \Delta_n)\mu_n(B_1, \Delta_n) \geq \mu_1(B_1, \Delta_n)^2 \geq C'n.$$

Two other applications of Minkowski's theorem on successive minima will be mentioned here. The first is a sharp form, due to Bombieri and Vaaler (1983), of 'Siegel's lemma'. In his investigations on transcendental numbers Siegel (1929) used Dirichlet's pigeonhole principle to prove that if  $A = (\alpha_{jk})$  is an  $m \times n$  matrix of integers, where

$m < n$ , such that  $|\alpha_{jk}| \leq \beta$  for all  $j, k$ , then the system of homogeneous linear equations

$$Ax = 0$$

has a solution  $x = (\xi_k)$  in integers, not all 0, such that  $|\xi_k| \leq 1 + (n\beta)^{m/(n-m)}$  for all  $k$ . Bombieri and Vaaler show that, if  $A$  has rank  $m$  and if  $g > 0$  is the greatest common divisor of all  $m \times m$  subdeterminants of  $A$ , then there are  $n - m$  linearly independent integral solutions  $x_j = (\xi_{jk})$  ( $j = 1, \dots, n - m$ ) such that

$$\prod_{j=1}^{n-m} \|x_j\| \leq [\det(AA^t)]^{1/2}/g,$$

where  $\|x_j\| = \max_k |\xi_{jk}|$ .

The second application, due to Gillet and Soulé (1991), may be regarded as an arithmetic analogue of the Riemann–Roch theorem for function fields. Again let  $K$  be a compact symmetric convex subset of  $\mathbb{R}^n$  with nonempty interior and let  $\mu_i$  denote the infimum of all  $\rho > 0$  such that  $\rho K$  contains at least  $i$  linearly independent points of  $\mathbb{Z}^n$  ( $i = 1, \dots, n$ ). If  $M(K)$  is the number of points of  $\mathbb{Z}^n$  in  $K$ , and if  $h$  is the maximum number of linearly independent points of  $\mathbb{Z}^n$  in the interior of  $K$ , then Gillet and Soulé show that  $\mu_1 \cdots \mu_h / M(K)$  is bounded above and below by positive constants, which depend on  $n$  but not on  $K$ .

A number of results in this section have dealt with compact symmetric convex sets with nonempty interior. Since such sets may appear rather special, it should be pointed out that they arise very naturally in connection with normed vector spaces.

The vector space  $\mathbb{R}^n$  is said to be *normed* if with each  $x \in \mathbb{R}^n$  there is associated a real number  $|x|$  with the properties

- (i)  $|x| \geq 0$ , with equality if and only if  $x = O$ ,
- (ii)  $|x + y| \leq |x| + |y|$  for all  $x, y \in \mathbb{R}^n$ ,
- (iii)  $|\alpha x| = |\alpha||x|$  for all  $x \in \mathbb{R}^n$  and all  $\alpha \in \mathbb{R}$ .

Let  $K$  denote the set of all  $x \in \mathbb{R}^n$  such that  $|x| \leq 1$ . Then  $K$  is bounded, since all norms on a finite-dimensional vector space are equivalent. In fact  $K$  is compact, since it follows from (ii) that  $K$  is closed. Moreover  $K$  is convex and symmetric, by (ii) and (iii). Furthermore, by (i) and (iii),  $x/|x| \in K$  for each nonzero  $x \in \mathbb{R}^n$ . Hence the interior of  $K$  is nonempty and is actually the set of all  $x \in \mathbb{R}^n$  such that  $|x| < 1$ .

Conversely, let  $K$  be a compact symmetric convex subset of  $\mathbb{R}^n$  with nonempty interior. Then the origin is an interior point of  $K$  and for each nonzero  $x \in \mathbb{R}^n$  there is a unique  $\rho > 0$  such that  $\rho x$  is on the boundary of  $K$ . If we put  $|x| = \rho^{-1}$ , and  $|O| = 0$ , then (i) obviously holds. Furthermore, since  $|-x| = |x|$ , it is easily seen that (iii) holds. Finally, if  $y \in \mathbb{R}^n$  and  $|y| = \sigma^{-1}$ , then  $\rho x, \sigma y \in K$  and hence, since  $K$  is convex,

$$\rho\sigma(\rho + \sigma)^{-1}(x + y) = \sigma(\rho + \sigma)^{-1}\rho x + \rho(\rho + \sigma)^{-1}\sigma y \in K.$$

Hence

$$|x + y| \leq (\rho + \sigma)/\rho\sigma = |x| + |y|.$$

Thus  $\mathbb{R}^n$  is a normed vector space and  $K$  the set of all  $x \in \mathbb{R}^n$  such that  $|x| \leq 1$ .

#### 4 Voronoi Cells

Throughout this section we suppose  $\mathbb{R}^n$  equipped with the *Euclidean metric*:

$$d(y, z) = \|y - z\|,$$

where  $\|x\| = (x^t x)^{1/2}$ . We call  $\|x\|^2 = x^t x$  the *square-norm* of  $x$  and we denote the scalar product  $y^t z$  by  $(y, z)$ .

Fix some point  $x_0 \in \mathbb{R}^n$ . For any point  $x \neq x_0$ , the set of all points which are equidistant from  $x_0$  and  $x$  is the hyperplane  $H_x$  which passes through the midpoint of the segment joining  $x_0$  and  $x$  and is orthogonal to this segment. Analytically,  $H_x$  is the set of all  $y \in \mathbb{R}^n$  such that

$$(x - x_0, y) = (x - x_0, x + x_0)/2,$$

which simplifies to

$$2(x - x_0, y) = \|x\|^2 - \|x_0\|^2.$$

The set of all points which are closer to  $x_0$  than to  $x$  is the open half-space  $G_x$  consisting of all points  $y \in \mathbb{R}^n$  such that

$$2(x - x_0, y) < \|x\|^2 - \|x_0\|^2.$$

The closed half-space  $\bar{G}_x = H_x \cup G_x$  is the set of all points at least as close to  $x_0$  as to  $x$ .

Let  $X$  be a subset of  $\mathbb{R}^n$  containing more than one point which is *discrete*, i.e. for each  $y \in \mathbb{R}^n$  there exists an open set containing  $y$  which contains at most one point of  $X$ . It follows that each bounded subset of  $\mathbb{R}^n$  contains only finitely many points of  $X$  since, if there were infinitely many, they would have an accumulation point. Hence for each  $y \in \mathbb{R}^n$  there exists an  $x_0 \in X$  whose distance from  $y$  is minimal:

$$d(x_0, y) \leq d(x, y) \quad \text{for every } x \in X. \quad (1)$$

For each  $x_0 \in X$  we define its *Voronoi cell*  $V(x_0)$  to be the set of all  $y \in \mathbb{R}^n$  for which (1) holds. Voronoi cells are also called ‘Dirichlet domains’, since they were used by Dirichlet (1850) in  $\mathbb{R}^2$  before Voronoi (1908) used them in  $\mathbb{R}^n$ .

If we choose  $r > 0$  so that the open ball

$$\beta_r(x_0) := \{y \in \mathbb{R}^n : d(x_0, y) < r\}$$

contains no point of  $X$  except  $x_0$ , then  $\beta_{r/2}(x_0) \subseteq V(x_0)$ . Thus  $x_0$  is an interior point of  $V(x_0)$ .

Since

$$\bar{G}_x = \{y \in \mathbb{R}^n : d(x_0, y) \leq d(x, y)\},$$

we have  $V(x_0) \subseteq \bar{G}_x$  and actually

$$V(x_0) = \bigcap_{x \in X \setminus \{x_0\}} \bar{G}_x. \quad (2)$$

It follows at once from (2) that  $V(x_0)$  is closed and convex. Hence  $V(x_0)$  is the closure of its nonempty interior.

According to the definitions of §3, the Voronoi cells form a tiling of  $\mathbb{R}^n$ , since

$$\mathbb{R}^n = \bigcup_{x \in X} V(x),$$

$$\text{int}V(x) \cap \text{int}V(x') = \emptyset \quad \text{if } x, x' \in X \text{ and } x \neq x'.$$

A subset  $A$  of a convex set  $C$  is said to be a *face* of  $C$  if  $A$  is convex and, for any  $c, c' \in C$ ,  $(c, c') \cap A \neq \emptyset$  implies  $c, c' \in A$ . The tiling by Voronoi cells has the additional property that  $V(x) \cap V(x')$  is a face of both  $V(x)$  and  $V(x')$  if  $x, x' \in X$  and  $x \neq x'$ . We will prove this by showing that if  $y_1, y_2$  are distinct points of  $V(x)$  and if  $z \in (y_1, y_2) \cap V(x')$ , then  $y_1 \in V(x')$ .

Since  $z \in V(x) \cap V(x')$ , we have  $d(x, z) = d(x', z)$ . Thus  $z$  lies on the hyperplane  $H$  which passes through the midpoint of the segment joining  $x$  and  $x'$  and is orthogonal to this segment. If  $y_1 \notin V(x')$ , then  $d(x, y_1) < d(x', y_1)$ . Thus  $y_1$  lies in the open half-space  $G$  associated with the hyperplane  $H$  which contains  $x$ . But then  $y_2$  lies in the open half-space  $G'$  which contains  $x'$ , i.e.  $d(x', y_2) < d(x, y_2)$ , which contradicts  $y_2 \in V(x)$ .

We now assume that the set  $X$  is not only discrete, but also *relatively dense*, i.e.

(†) there exists  $R > 0$  such that, for each  $y \in \mathbb{R}^n$ , there is some  $x \in X$  with  $d(x, y) < R$ .

It follows at once that  $V(x_0) \subseteq \beta_R(x_0)$ . Thus  $V(x_0)$  is bounded and, since it is closed, even compact. The ball  $\beta_{2R}(x_0)$  contains only finitely many points  $x_1, \dots, x_m$  of  $X$  apart from  $x_0$ . We are going to show that

$$V(x_0) = \bigcap_{i=1}^m \bar{G}_{x_i}. \quad (3)$$

By (2) we need only show that if  $y \in \bigcap_{i=1}^m \bar{G}_{x_i}$ , then  $y \in \bar{G}_x$  for every  $x \in X$ .

Assume that  $d(x_0, y) \geq R$  and choose  $z$  on the segment joining  $x_0$  and  $y$  so that  $d(x_0, z) = R$ . For some  $x \in X$  we have  $d(x, z) < R$  and hence  $0 < d(x_0, x) < 2R$ . Consequently  $x = x_i$  for some  $i \in \{1, \dots, m\}$ . Since  $d(x_i, z) < R = d(x_0, z)$ , we have  $z \notin \bar{G}_{x_i}$ . But this is a contradiction, since  $x_0, y \in \bar{G}_{x_i}$  and  $z$  is on the segment joining them.

We conclude that  $d(x_0, y) < R$ . If  $x \in X$  and  $x \neq x_0, x_1, \dots, x_m$ , then

$$\begin{aligned} d(x, y) &\geq d(x_0, x) - d(x_0, y) \\ &\geq 2R - R = R > d(x_0, y). \end{aligned}$$

Consequently  $y \in \bar{G}_x$  for every  $x \in X$ .

It follows from (3) that  $V(x_0)$  is a polyhedron. Since  $V(x_0)$  is bounded and has a nonempty interior, it is actually an *n-dimensional polytope*.

The faces of a polytope are an important part of its structure. An  $(n-1)$ -dimensional face of an  $n$ -dimensional polytope is said to be a *facet* and a 0-dimensional face is said to be a *vertex*. We now apply to  $V(x_0)$  some properties common to all polytopes.

In the representation (3) it may be possible to omit some closed half-spaces  $\bar{G}_{x_i}$  without affecting the validity of the representation. By omitting as many half-spaces as possible we obtain an *irredundant representation*, which by suitable choice of notation we may take to be

$$V(x_0) = \bigcap_{i=1}^l \bar{G}_{x_i}$$

for some  $l \leq m$ . The intersections  $V(x_0) \cap H_{x_i}$  ( $1 \leq i \leq l$ ) are then the distinct facets of  $V(x_0)$ . Any nonempty proper face of  $V(x_0)$  is contained in a facet and is the intersection of those facets which contain it. Furthermore, any nonempty face of  $V(x_0)$  is the convex hull of those vertices of  $V(x_0)$  which it contains.

It follows that for each  $x_i$  ( $1 \leq i \leq l$ ) there is a vertex  $v_i$  of  $V(x_0)$  such that

$$d(x_0, v_i) = d(x_i, v_i).$$

For  $d(x_0, v) \leq d(x_i, v)$  for every vertex  $v$  of  $V(x_0)$ . Assume that  $d(x_0, v) < d(x_i, v)$  for every vertex  $v$  of  $V(x_0)$ . Then the open half-space  $G_{x_i}$  contains all vertices  $v$  and hence also their convex hull  $V(x_0)$ . But this is a contradiction, since  $V(x_0) \cap H_{x_i}$  is a facet of  $V(x_0)$ .

To illustrate these results take  $X = \mathbb{Z}^n$  and  $x_0 = O$ . Then the Voronoi cell  $V(O)$  is the cube consisting of all points  $y = (\eta_1, \dots, \eta_n) \in \mathbb{R}^n$  with  $|\eta_i| \leq 1/2$  ( $i = 1, \dots, n$ ). It has the minimal number  $2n$  of facets.

In fact any lattice  $A$  in  $\mathbb{R}^n$  is discrete and has the property ( $\dagger$ ). For a lattice  $A$  we can restrict attention to the Voronoi cell  $V(A) := V(O)$ , since an arbitrary Voronoi cell is obtained from it by a translation:  $V(x_0) = V(O) + x_0$ . The Voronoi cell of a lattice has extra properties. Since  $x \in A$  implies  $-x \in A$ ,  $y \in V(A)$  implies  $-y \in V(A)$ . Furthermore, if  $x_i$  is a lattice vector determining a facet of  $V(A)$  and if  $y \in V(A) \cap H_{x_i}$ , then  $\|y\| = \|y - x_i\|$ . Since  $x \in A$  implies  $x_i - x \in A$ , it follows that  $y \in V(A) \cap H_{x_i}$  implies  $x_i - y \in V(A) \cap H_{x_i}$ . Thus the Voronoi cell  $V(A)$  and all its facets are centrosymmetric.

In addition, any orthogonal transformation of  $\mathbb{R}^n$  which maps onto itself the lattice  $A$  also maps onto itself the Voronoi cell  $V(A)$ . Furthermore the Voronoi cell  $V(A)$  has volume  $d(A)$ , by Proposition 11, since the lattice translates of  $V(A)$  form a tiling of  $\mathbb{R}^n$ .

We define a *facet vector* or 'relevant vector' of a lattice  $A$  to be a vector  $x_i \in A$  such that  $V(A) \cap H_{x_i}$  is a facet of the Voronoi cell  $V(A)$ . If  $V(A)$  is contained in the closed ball  $B_R = \{x \in \mathbb{R}^n : \|x\| \leq R\}$ , then every facet vector  $x_i$  satisfies  $\|x_i\| \leq 2R$ . For, if  $y \in V(A) \cap H_{x_i}$  then, by Schwarz's inequality (Chapter I, §4),

$$\|x_i\|^2 = 2(x_i, y) \leq 2\|x_i\|\|y\|.$$

The facet vectors were characterized by Voronoi (1908) in the following way:

**Proposition 13** A nonzero vector  $x \in A$  is a facet vector of the lattice  $A \subseteq \mathbb{R}^n$  if and only if every vector  $x' \in x + 2A$ , except  $\pm x$ , satisfies  $\|x'\| > \|x\|$ .

*Proof* Suppose first that  $\|x\| < \|x'\|$  for all  $x' \neq \pm x$  such that  $(x' - x)/2 \in A$ . If  $z \in A$  and  $x' = 2z - x$ , then  $(x' - x)/2 \in A$ . Hence if  $z \neq O, x$  then

$$\|x/2\| < \|z - x/2\|,$$

i.e.  $x/2 \in G_z$ . Since  $\|x/2\| = \|x - x/2\|$ , it follows that  $x/2 \in V(A)$  and  $x$  is a facet vector.

Suppose next that there exists  $x' \neq \pm x$  such that  $w = (x' - x)/2 \in A$  and  $\|x'\| \leq \|x\|$ . Then also  $z = (x' + x)/2 \in A$  and  $z, w \neq O$ . If  $y \in \bar{G}_z \cap \bar{G}_{-w}$ , then

$$2(z, y) \leq \|z\|^2, \quad -2(w, y) \leq \|w\|^2.$$

Hence, by the parallelogram law (Chapter I, §10),

$$\begin{aligned} 2(x, y) &= 2(z, y) - 2(w, y) \leq \|z\|^2 + \|w\|^2 \\ &= \|x\|^2/2 + \|x'\|^2/2 \leq \|x\|^2. \end{aligned}$$

That is,  $y \in \bar{G}_x$ . Thus  $\bar{G}_x$  is not needed to define  $V(A)$  and  $x$  is not a facet vector.  $\square$

Any lattice  $A$  contains a nonzero vector with minimal square-norm. Such a vector will be called a *minimal vector*. Its square-norm will be called the *minimum* of  $A$  and will be denoted by  $m(A)$ .

**Proposition 14** *If  $A \subseteq \mathbb{R}^n$  is a lattice with minimum  $m(A)$ , then any nonzero vector in  $A$  with square-norm  $< 2m(A)$  is a facet vector. In particular, any minimal vector is a facet vector.*

*Proof* Put  $r = m(A)$  and let  $x$  be a nonzero vector in  $A$  with  $\|x\|^2 < 2r$ . If  $x$  is not a facet vector, there exists  $y \neq \pm x$  with  $(y - x)/2 \in A$  such that  $\|y\| \leq \|x\|$ . Since  $(y \pm x)/2 \in A$ ,  $\|x \pm y\|^2 \geq 4r$ . Thus

$$4r \leq \|x\|^2 + \|y\|^2 \pm 2(x, y) < 4r \pm 2(x, y),$$

which is impossible.  $\square$

**Proposition 15** *For any lattice  $A \subseteq \mathbb{R}^n$ , the number of facets of its Voronoi cell  $V(A)$  is at most  $2(2^n - 1)$ .*

*Proof* Let  $x_1, \dots, x_n$  be a basis for  $A$ . Then any vector  $x \in A$  has a unique representation  $x = x' + x''$ , where  $x' \in 2A$  and

$$x'' = \alpha_1 x_1 + \dots + \alpha_n x_n,$$

with  $\alpha_j \in \{0, 1\}$  for  $j = 1, \dots, n$ . Thus the number of cosets of  $2A$  in  $A$  is  $2^n$ . But, by Proposition 13, each coset contains at most one pair  $\pm y$  of facet vectors. Since  $2A$  itself does not contain any facet vectors, the total number of facet vectors is at most  $2(2^n - 1)$ .  $\square$

There exist lattices  $A \subseteq \mathbb{R}^n$  for which the upper bound of Proposition 15 is attained, e.g. the lattice  $A = \{Tz : z \in \mathbb{Z}^n\}$  with  $T = I + \beta J$ , where  $J$  denotes the  $n \times n$  matrix every element of which is 1 and  $\beta = \{(1 + n)^{1/2} - 1\}/n$ .

**Proposition 16** *Every vector of a lattice  $A \subseteq \mathbb{R}^n$  is an integral linear combination of facet vectors.*

*Proof* Let  $b_1, \dots, b_m$  be the facet vectors of  $A$  and put

$$A' = \{x = \beta_1 b_1 + \dots + \beta_m b_m : \beta_1, \dots, \beta_m \in \mathbb{Z}\}.$$

Evidently  $A'$  is a subgroup of  $\mathbb{R}^n$  and actually a discrete subgroup, since  $A' \subseteq A$ . If  $A'$  were contained in a hyperplane of  $\mathbb{R}^n$  any point on the line through the origin orthogonal to this hyperplane would belong to the Voronoi cell  $V$  of  $A$ , which is impossible because  $V$  is bounded. Hence  $A'$  contains  $n$  linearly independent vectors.

Thus  $A'$  is a sublattice of  $A$ . It follows that the Voronoi cell  $V$  of  $A$  is contained in the Voronoi cell  $V'$  of  $A'$ . But if  $y \in V'$ , then

$$\|y\| \leq \|b_i - y\|, \quad (i = 1, \dots, m)$$

and hence  $y \in V$ . Thus  $V' = V$ . Hence the  $A'$ -translates of  $V$  and the  $A$ -translates of  $V$  are both tilings of  $\mathbb{R}^n$ . Since  $A' \subseteq A$ , this is possible only if  $A' = A$ .  $\square$

Since every integral linear combination of facet vectors is in the lattice, Proposition 16 implies

**Corollary 17** *Distinct lattices in  $\mathbb{R}^n$  have distinct Voronoi cells.*

Proposition 16 does not say that the lattice has a basis of facet vectors. It is known that every lattice in  $\mathbb{R}^n$  has a basis of facet vectors if  $n \leq 6$ , but if  $n > 6$  this is still an open question. It is known also that every lattice in  $\mathbb{R}^n$  has a basis of minimal vectors when  $n \leq 4$  but, when  $n > 4$ , there are lattices with no such basis. In fact a lattice may have no basis of minimal vectors, even though every lattice vector is an integral linear combination of minimal vectors.

Lattices and their Voronoi cells have long been used in crystallography. An  $n$ -dimensional *crystal* may be defined mathematically to be a subset of  $\mathbb{R}^n$  of the form

$$F + A = \{x + y : x \in F, y \in A\},$$

where  $F$  is a finite set and  $A$  a lattice. Crystals may be studied by means of their symmetry groups.

An *isometry* of  $\mathbb{R}^n$  is an invertible affine transformation which leaves unaltered the Euclidean distance between any two points. For example, any orthogonal transformation is an isometry and so is a translation by an arbitrary vector  $v$ . Any isometry is the composite of a translation and an orthogonal transformation. The *symmetry group* of a set  $X \subseteq \mathbb{R}^n$  is the group of all isometries of  $\mathbb{R}^n$  which map  $X$  to itself.

We define an  $n$ -dimensional *crystallographic group* to be a group  $G$  of isometries of  $\mathbb{R}^n$  such that the vectors corresponding to translations in  $G$  form an  $n$ -dimensional lattice. It is not difficult to show that a subset of  $\mathbb{R}^n$  is an  $n$ -dimensional crystal if and only if it is discrete and its symmetry group is an  $n$ -dimensional crystallographic group.

It was shown by Bieberbach (1911) that a group  $G$  of isometries of  $\mathbb{R}^n$  is a crystallographic group if and only if it is discrete and has a compact fundamental domain  $D$ , i.e. the sets  $\{g(D) : g \in G\}$  form a tiling of  $\mathbb{R}^n$ . He could then show that the translations in a crystallographic group form a torsion-free abelian normal subgroup of finite index. He showed later (1912) that two crystallographic groups  $G_1, G_2$  are isomorphic if and only if there exists an invertible affine transformation  $A$  such that

$G_2 = A^{-1}G_1A$ . With the aid of results of Minkowski and Jordan it follows that, for a given dimension  $n$ , there are only finitely many non-isomorphic crystallographic groups. These results provided a positive answer to the first part of the 18th Problem of Hilbert (1900).

The structure of physical crystals is analysed by means of the corresponding 3-dimensional crystallographic groups. A stronger concept than isomorphism is useful for such applications. Two crystallographic groups  $G_1, G_2$  may be said to be *properly isomorphic* if there exists an orientation-preserving invertible affine transformation  $A$  such that  $G_2 = A^{-1}G_1A$ . An isomorphism class of crystallographic groups either coincides with a proper isomorphism class or splits into two distinct proper isomorphism classes.

Fedorov (1891) showed that there are 17 isomorphism classes of 2-dimensional crystallographic groups, none of which splits. Collating earlier work of Sohncke (1879), Schoenflies (1889) and himself, Fedorov (1892) also showed that there are 219 isomorphism classes of 3-dimensional crystallographic groups, 11 of which split. More recently, Brown *et al.* (1978) have shown that there are 4783 isomorphism classes of 4-dimensional crystallographic groups, 112 of which split.

## 5 Densest Packings

The result of Hermite, mentioned at the beginning of the chapter, can be formulated in terms of lattices instead of quadratic forms. For any real non-singular matrix  $T$ , the matrix

$$A = T^t T$$

is a real positive definite symmetric matrix. Conversely, by a principal axes transformation, or more simply by induction, it may be seen that any real positive definite symmetric matrix  $A$  may be represented in this way.

Let  $\Lambda$  be the lattice

$$\Lambda = \{y = Tx \in \mathbb{R}^n : x \in \mathbb{Z}^n\}$$

and put

$$\gamma(A) = m(\Lambda)/d(\Lambda)^{2/n},$$

where  $d(\Lambda)$  is the determinant and  $m(\Lambda)$  the minimum of  $\Lambda$ . Then  $\gamma(\rho\Lambda) = \gamma(\Lambda)$  for any  $\rho > 0$ . Hermite's result that there exists a positive constant  $c_n$ , depending only on  $n$ , such that  $0 < x^t A x \leq c_n (\det A)^{1/n}$  for some  $x \in \mathbb{Z}^n$  may be restated in the form

$$\gamma(A) \leq c_n.$$

*Hermite's constant*  $\gamma_n$  is defined to be the least positive constant  $c_n$  such that this inequality holds for all  $\Lambda \subseteq \mathbb{R}^n$ .

It may be shown that  $\gamma_n^n$  is a rational number for each  $n$ . It follows from Proposition 2 that  $\lim_{n \rightarrow \infty} \gamma_n/n \leq 2/\pi e$ . Minkowski (1905) showed also that

$$\lim_{n \rightarrow \infty} \gamma_n/n \geq 1/2\pi e = 0.0585 \dots,$$

and it is possible that actually  $\lim_{n \rightarrow \infty} \gamma_n/n = 1/2\pi e$ . The significance of Hermite's constant derives from its connection with lattice packings of balls, as we now explain.

Let  $A$  be a lattice in  $\mathbb{R}^n$  and  $K$  a subset of  $\mathbb{R}^n$  which is the closure of a nonempty open set  $G$ . We say that  $A$  gives a *lattice packing* for  $K$  if the family of translates  $K+x$  ( $x \in A$ ) is a packing of  $\mathbb{R}^n$ , i.e. if for any two distinct points  $x, y \in A$  the interiors  $G+x$  and  $G+y$  are disjoint. This is the same as saying that  $A$  does not contain the difference of any two distinct points of the interior of  $K$ , since  $g+x = g'+y$  if and only if  $g'-g = x-y$ . If  $K$  is a compact symmetric convex set with nonempty interior  $G$ , it is the same as saying that the interior of the set  $2K$  contains no nonzero point of  $A$ , since in this case  $g, g' \in G$  implies  $(g'-g)/2 \in G$  and  $2g = g - (-g)$ .

The *density* of the lattice packing, i.e. the fraction of the total space which is occupied by translates of  $K$ , is clearly  $\lambda(K)/d(A)$ . Hence the maximum density of any lattice packing for  $K$  is

$$\delta(K) = \lambda(K)/\Delta(2K) = 2^{-n}\lambda(K)/\Delta(K),$$

where  $\Delta(K)$  is the critical determinant of  $K$ , as defined in §3. The use of the word 'maximum' is justified, since it will be shown in §6 that the infimum involved in the definition of critical determinant is attained.

Our interest is in the special case of a closed ball:  $K = B_\rho = \{x \in \mathbb{R}^n : \|x\| \leq \rho\}$ . By what we have said,  $A$  gives a lattice packing for  $B_\rho$  if and only if the interior of  $B_{2\rho}$  contains no nonzero point of  $A$ , i.e. if and only if  $m(A)^{1/2} \geq 2\rho$ . Hence

$$\begin{aligned} \delta(B_\rho) &= \sup\{\lambda(B_\rho)/d(A) : m(A)^{1/2} = 2\rho\} \\ &= \kappa_n \rho^n \sup\{d(A)^{-1} : m(A)^{1/2} = 2\rho\}, \end{aligned}$$

where  $\kappa_n = \pi^{n/2}/(n/2)!$  again denotes the volume of the unit ball in  $\mathbb{R}^n$ . By virtue of homogeneity it follows that

$$\delta_n := \delta(B_\rho) = 2^{-n}\kappa_n \sup_A \gamma(A)^{n/2},$$

where the supremum is now over all lattices  $A \subseteq \mathbb{R}^n$ ; that is, in terms of Hermite's constant  $\gamma_n$ ,

$$\delta_n = 2^{-n}\kappa_n \gamma_n^{n/2}.$$

Thus  $\gamma_n$ , like  $\delta_n$ , measures the densest lattice packing of balls. A lattice  $A \subseteq \mathbb{R}^n$  for which  $\gamma(A) = \gamma_n$ , i.e. a critical lattice for a ball, will be called simply a *densest lattice*.

The densest lattice in  $\mathbb{R}^n$  is known for each  $n \leq 8$ , and is uniquely determined apart from isometries and scalar multiples. In fact these densest lattices are all examples of indecomposable root lattices. These terms will now be defined.

A lattice  $A$  is said to be *decomposable* if there exist additive subgroups  $A_1, A_2$  of  $A$ , each containing a nonzero vector, such that  $(x_1, x_2) = 0$  for all  $x_1 \in A_1$  and  $x_2 \in A_2$ , and every vector in  $A$  is the sum of a vector in  $A_1$  and a vector in  $A_2$ . Since  $A_1$  and  $A_2$  are necessarily discrete, they are lattices in the wide sense (i.e. they are not

full-dimensional). We say also that  $A$  is the *orthogonal sum* of the lattices  $A_1$  and  $A_2$ . The orthogonal sum of any finite number of lattices is defined similarly. A lattice is *indecomposable* if it is not decomposable.

The following result was first proved by Eichler (1952).

**Proposition 18** *Any lattice  $A$  is an orthogonal sum of finitely many indecomposable lattices, which are uniquely determined apart from order.*

*Proof* (i) Define a vector  $x \in A$  to be ‘decomposable’ if there exist nonzero vectors  $x_1, x_2 \in A$  such that  $x = x_1 + x_2$  and  $(x_1, x_2) = 0$ . We show first that every nonzero  $x \in A$  is a sum of finitely many indecomposable vectors.

By definition,  $x$  is either indecomposable or is the sum of two nonzero orthogonal vectors in  $A$ . Both these vectors have square-norm less than the square-norm of  $x$ , and for each of them the same alternative presents itself. Continuing in this way, we must eventually arrive at indecomposable vectors, since there are only finitely many vectors in  $A$  with square-norm less than that of  $x$ .

(ii) If  $A$  is the orthogonal sum of finitely many lattices  $L_v$  then, by the definition of an orthogonal sum, every indecomposable vector of  $A$  lies in one of the sublattices  $L_v$ . Hence if two indecomposable vectors are not orthogonal, they lie in the same sublattice  $L_v$ .

(iii) Call two indecomposable vectors  $x, x'$  ‘equivalent’ if there exist indecomposable vectors  $x = x_0, x_1, \dots, x_{k-1}, x_k = x'$  such that  $(x_j, x_{j+1}) \neq 0$  for  $0 \leq j < k$ . Clearly ‘equivalence’ is indeed an equivalence relation and thus the set of all indecomposable vectors is partitioned into equivalence classes  $\mathcal{C}_\mu$ . Two vectors from different equivalence classes are orthogonal and, if  $A$  is an orthogonal sum of lattices  $L_v$  as in (ii), then two vectors from the same equivalence class lie in the same sublattice  $L_v$ .

(iv) Let  $A_\mu$  be the subgroup of  $A$  generated by the vectors in the equivalence class  $\mathcal{C}_\mu$ . Then, by (i),  $A$  is generated by the sublattices  $A_\mu$ . Since, by (iii),  $A_\mu$  is orthogonal to  $A_{\mu'}$  if  $\mu \neq \mu'$ ,  $A$  is actually the orthogonal sum of the sublattices  $A_\mu$ . If  $A$  is an orthogonal sum of lattices  $L_v$  as in (ii), then each  $A_\mu$  is contained in some  $L_v$ . It follows that each  $A_\mu$  is indecomposable and that these indecomposable sublattices are uniquely determined apart from order.  $\square$

Let  $A$  be a lattice in  $\mathbb{R}^n$ . If  $A \subseteq A^*$ , i.e. if  $(x, y) \in \mathbb{Z}$  for all  $x, y \in A$ , then  $A$  is said to be *integral*. If  $(x, x)$  is an even integer for every  $x \in A$ , then  $A$  is said to be *even*. (It follows that an even lattice is also integral.) If  $A$  is even and every vector in  $A$  is an integral linear combination of vectors in  $A$  with square-norm 2, then  $A$  is said to be a *root lattice*.

Thus in a root lattice the minimal vectors have square-norm 2. It may be shown by a long, but elementary, argument that any root lattice has a basis of minimal vectors such that every minimal vector is an integral linear combination of the basis vectors with coefficients which are all nonnegative or all nonpositive. Such a basis will be called a *simple basis*. The facet vectors of a root lattice are precisely the minimal vectors, and hence its Voronoi cell is the set of all  $y \in \mathbb{R}^n$  such that  $(y, x) \leq 1$  for every minimal vector  $x$ .

Any root lattice is an orthogonal sum of indecomposable root lattices. It was shown by Witt (1941) that the indecomposable root lattices can be completely enumerated:

Table 1. Indecomposable root lattices

$A_n = \{x = (\zeta_0, \zeta_1, \dots, \zeta_n) \in \mathbb{Z}^{n+1} : \zeta_0 + \zeta_1 + \dots + \zeta_n = 0\} \ (n \geq 1);$
$D_n = \{x = (\zeta_1, \dots, \zeta_n) \in \mathbb{Z}^n : \zeta_1 + \dots + \zeta_n \text{ even}\} \ (n \geq 3);$
$E_8 = D_8 \cup D_8^\dagger$ , where $D_8^\dagger = (1/2, 1/2, \dots, 1/2) + D_8$ ;
$E_7 = \{x = (\zeta_1, \dots, \zeta_8) \in E_8 : \zeta_7 = -\zeta_8\};$
$E_6 = \{x = (\zeta_1, \dots, \zeta_8) \in E_8 : \zeta_6 = \zeta_7 = -\zeta_8\}.$

they are all listed in Table 1. We give also their minimal vectors in terms of the canonical basis  $e_1, \dots, e_n$  of  $\mathbb{R}^n$ .

The lattice  $A_n$  has  $n(n + 1)$  minimal vectors, namely the vectors  $\pm(e_j - e_k)$  ( $0 \leq j < k \leq n$ ), and the vectors  $e_0 - e_1, e_1 - e_2, \dots, e_{n-1} - e_n$  form a simple basis. By calculating the determinant of  $B^t B$ , where  $B$  is the  $(n + 1) \times n$  matrix whose columns are the vectors of this simple basis, it may be seen that the determinant of the lattice  $A_n$  is  $(n + 1)^{1/2}$ .

The lattice  $D_n$  has  $2n(n - 1)$  minimal vectors, namely the vectors  $\pm e_j \pm e_k$  ( $1 \leq j < k \leq n$ ). The vectors  $e_1 - e_2, e_2 - e_3, \dots, e_{n-1} - e_n, e_{n-1} + e_n$  form a simple basis and hence the lattice  $D_n$  has determinant 2.

The lattice  $E_8$  has 240 minimal vectors, namely the 112 vectors  $\pm e_j \pm e_k$  ( $1 \leq j < k \leq 8$ ) and the 128 vectors  $(\pm e_1 \pm \dots \pm e_8)/2$  with an even number of minus signs. The vectors

$$\begin{aligned} v_1 &= (e_1 - e_2 - \dots - e_7 + e_8)/2, & v_2 &= e_1 + e_2, \\ v_3 &= e_2 - e_1, & v_4 &= e_3 - e_2, \dots, & v_8 &= e_7 - e_6, \end{aligned}$$

form a simple basis and hence the lattice has determinant 1.

The lattice  $E_7$  has 126 minimal vectors, namely the 60 vectors  $\pm e_j \pm e_k$  ( $1 \leq j < k \leq 6$ ), the vectors  $\pm(e_7 - e_8)$  and the 64 vectors  $\pm\left(\sum_{i=1}^6 (\pm e_i) - e_7 + e_8\right)/2$  with an odd number of minus signs in the sum. The vectors  $v_1, \dots, v_7$  form a simple basis and the lattice has determinant  $\sqrt{2}$ .

The lattice  $E_6$  has 72 minimal vectors, namely the 40 vectors  $\pm e_j \pm e_k$  ( $1 \leq j < k \leq 5$ ) and the 32 vectors  $\pm\left(\sum_{i=1}^5 (\pm e_i) - e_6 - e_7 + e_8\right)/2$  with an even number of minus signs in the sum. The vectors  $v_1, \dots, v_6$  form a simple basis and the lattice has determinant  $\sqrt{3}$ .

We now return to lattice packings of balls. The densest lattices for  $n \leq 8$  are given in Table 2. These lattices were shown to be densest by Lagrange (1773) for  $n = 2$ , by Gauss (1831) for  $n = 3$ , by Korkine and Zolotareff (1872, 1877) for  $n = 4, 5$  and by Blichfeldt (1925, 1926, 1934) for  $n = 6, 7, 8$ .

Although the densest lattice in  $\mathbb{R}^n$  is unknown for every  $n > 8$ , there are plausible candidates in some dimensions. In particular, a lattice discovered by Leech (1967) is believed to be densest in 24 dimensions. This lattice may be constructed in the following way. Let  $p$  be a prime such that  $p \equiv 3 \pmod{4}$  and let  $H_n$  be the Hadamard matrix of order  $n = p + 1$  constructed by Paley's method (See Chapter V, §2). The columns of the matrix

Table 2. Densest lattices in  $\mathbb{R}^n$

$n$	$A$	$\gamma_n$	$\delta_n$
1	$A_1$	1	1
2	$A_2$	$(4/3)^{1/2} = 1.1547 \dots$	$3^{1/2}\pi/6 = 0.9068 \dots$
3	$D_3$	$2^{1/3} = 1.2599 \dots$	$2^{1/2}\pi/6 = 0.7404 \dots$
4	$D_4$	$2^{1/2} = 1.4142 \dots$	$\pi^2/16 = 0.6168 \dots$
5	$D_5$	$8^{1/5} = 1.5157 \dots$	$2^{1/2}\pi^2/30 = 0.4652 \dots$
6	$E_6$	$(64/3)^{1/6} = 1.6653 \dots$	$3^{1/2}\pi^3/144 = 0.3729 \dots$
7	$E_7$	$(64)^{1/7} = 1.8114 \dots$	$\pi^3/105 = 0.2952 \dots$
8	$E_8$	2	$\pi^4/384 = 0.2536 \dots$

$$T = (n/4 + 1)^{-1/2} \begin{pmatrix} (n/4 + 1)I_n & H_n - I_n \\ 0_n & I_n \end{pmatrix}$$

generate a lattice in  $\mathbb{R}^{2n}$ . For  $p = 3$  we obtain the root lattice  $E_8$  and for  $p = 11$  the Leech lattice  $A_{24}$ .

Leech’s lattice may be characterized as the unique even lattice  $A$  in  $\mathbb{R}^{24}$  with  $d(A) = 1$  and  $m(A) > 2$ . It was shown by Conway (1969) that, if  $G$  is the group of all orthogonal transformations of  $\mathbb{R}^{24}$  which map the Leech lattice  $A_{24}$  onto itself, then the factor group  $G/\{\pm I_{24}\}$  is a finite simple group, and two more finite simple groups are easily obtained as (stabilizer) subgroups. These are three of the 26 sporadic simple groups which were mentioned in §7 of Chapter V.

Leech’s lattice has 196560 minimal vectors of square-norm 4. Thus the packing of unit balls associated with  $A_{24}$  is such that each ball touches 196560 other balls. It has been shown that 196560 is the maximal number of nonoverlapping unit balls in  $\mathbb{R}^{24}$  which can touch another unit ball and that, up to isometry, there is only one possible arrangement.

Similarly, since  $E_8$  has 240 minimal vectors of square-norm 2, the packing of balls of radius  $2^{-1/2}$  associated with  $E_8$  is such that each ball touches 240 other balls. It has been shown that 240 is the maximal number of nonoverlapping balls of fixed radius in  $\mathbb{R}^8$  which can touch another ball of the same radius and that, up to isometry, there is only one possible arrangement.

In general, one may ask what is the *kissing number* of  $\mathbb{R}^n$ , i.e. the maximal number of nonoverlapping unit balls in  $\mathbb{R}^n$  which can touch another unit ball? The question, for  $n = 3$ , first arose in 1694 in a discussion between Newton, who claimed that the answer was 12, and Gregory, who said 13. It was first shown by Hoppe (1874) that Newton was right, but in this case the arrangement of the 12 balls in  $\mathbb{R}^3$  is *not* unique up to isometry. One possibility is to take the centres of the 12 balls to be the vertices of a regular icosahedron, the centre of which is the centre of the unit ball they touch.

The kissing number of  $\mathbb{R}^1$  is clearly 2. It is not difficult to show that the kissing number of  $\mathbb{R}^2$  is 6 and that the centres of the six unit balls must be the vertices of a regular hexagon, the centre of which is the centre of the unit ball they touch. For  $n > 3$  the kissing number of  $\mathbb{R}^n$  is unknown, except for the two cases  $n = 8$  and  $n = 24$  already mentioned.

## 6 Mahler's Compactness Theorem

It is useful to study not only individual lattices, but also the family  $\mathcal{L}_n$  of all lattices in  $\mathbb{R}^n$ . A sequence of lattices  $A_k \in \mathcal{L}_n$  will be said to *converge* to a lattice  $A \in \mathcal{L}_n$ , in symbols  $A_k \rightarrow A$ , if there exist bases  $b_{k1}, \dots, b_{kn}$  of  $A_k$  ( $k = 1, 2, \dots$ ) and a basis  $b_1, \dots, b_n$  of  $A$  such that

$$b_{kj} \rightarrow b_j \text{ as } k \rightarrow \infty \quad (j = 1, \dots, n).$$

Evidently this implies that  $d(A_k) \rightarrow d(A)$  as  $k \rightarrow \infty$ . Also, for any  $x \in A$  there exist  $x_k \in A_k$  such that  $x_k \rightarrow x$  as  $k \rightarrow \infty$ . In fact if  $x = \alpha_1 b_1 + \dots + \alpha_n b_n$ , where  $\alpha_i \in \mathbb{Z}$  ( $i = 1, \dots, n$ ), we can take  $x_k = \alpha_1 b_{k1} + \dots + \alpha_n b_{kn}$ .

It is not obvious from the definition that the limit of a sequence of lattices is uniquely determined, but this follows at once from the next result.

**Proposition 19** *Let  $A$  be a lattice in  $\mathbb{R}^n$  and let  $\{A_k\}$  be a sequence of lattices in  $\mathbb{R}^n$  such that  $A_k \rightarrow A$  as  $k \rightarrow \infty$ . If  $x_k \in A_k$  and  $x_k \rightarrow x$  as  $k \rightarrow \infty$ , then  $x \in A$ .*

*Proof* With the above notation,

$$x = \alpha_1 b_1 + \dots + \alpha_n b_n,$$

where  $\alpha_i \in \mathbb{R}$  ( $i = 1, \dots, n$ ), and similarly

$$x_k = \alpha_{k1} b_{k1} + \dots + \alpha_{kn} b_{kn},$$

where  $\alpha_{ki} \in \mathbb{R}$  and  $\alpha_{ki} \rightarrow \alpha_i$  as  $k \rightarrow \infty$  ( $i = 1, \dots, n$ ).

The linear transformation  $T_k$  of  $\mathbb{R}^n$  which maps  $b_i$  to  $b_{ki}$  ( $i = 1, \dots, n$ ) can be written in the form

$$T_k = I - A_k,$$

where  $A_k \rightarrow O$  as  $k \rightarrow \infty$ . It follows that

$$T_k^{-1} = (I - A_k)^{-1} = I + A_k + A_k^2 + \dots = I + C_k,$$

where also  $C_k \rightarrow O$  as  $k \rightarrow \infty$ . Hence

$$\begin{aligned} x_k &= T_k^{-1}(\alpha_{k1} b_{k1} + \dots + \alpha_{kn} b_{kn}) \\ &= (\alpha_{k1} + \eta_{k1}) b_{k1} + \dots + (\alpha_{kn} + \eta_{kn}) b_{kn}, \end{aligned}$$

where  $\eta_{ki} \rightarrow 0$  as  $k \rightarrow \infty$  ( $i = 1, \dots, n$ ). But  $\alpha_{ki} + \eta_{ki} \in \mathbb{Z}$  for every  $k$ . Letting  $k \rightarrow \infty$ , we obtain  $\alpha_i \in \mathbb{Z}$ . That is,  $x \in A$ .  $\square$

It is natural to ask if the Voronoi cells of a convergent sequence of lattices also converge in some sense. The required notion of convergence is in fact older than the notion of convergence of lattices and applies to arbitrary compact subsets of  $\mathbb{R}^n$ .

The *Hausdorff distance*  $h(K, K')$  between two compact subsets  $K, K'$  of  $\mathbb{R}^n$  is defined to be the infimum of all  $\rho > 0$  such that every point of  $K$  is distant at most  $\rho$  from some point of  $K'$  and every point of  $K'$  is distant at most  $\rho$  from some point

of  $K$ . We will show that this defines a metric, the *Hausdorff metric*, on the space of all compact subsets of  $\mathbb{R}^n$ .

Evidently

$$0 \leq h(K, K') = h(K', K) < \infty.$$

Moreover  $h(K, K') = 0$  implies  $K = K'$ . For if  $x' \in K'$ , there exist  $x_k \in K$  such that  $x_k \rightarrow x'$  and hence  $x' \in K$ , since  $K$  is closed. Thus  $K' \subseteq K$ , and similarly  $K \subseteq K'$ .

Finally we prove the triangle inequality

$$h(K, K'') \leq h(K, K') + h(K', K'').$$

To simplify writing, put  $\rho = h(K, K')$  and  $\rho' = h(K', K'')$ . For any  $\varepsilon > 0$ , if  $x \in K$  there exist  $x' \in K'$  such that  $\|x - x'\| < \rho + \varepsilon$  and then  $x'' \in K''$  such that  $\|x' - x''\| < \rho' + \varepsilon$ . Hence

$$\|x - x''\| < \rho + \rho' + 2\varepsilon.$$

Similarly, if  $x'' \in K''$  there exists  $x \in K$  for which the same inequality holds. Since  $\varepsilon$  can be arbitrarily small, this completes the proof.

The definition of Hausdorff distance can also be expressed in the form

$$h(K, K') = \inf\{\rho \geq 0 : K \subseteq K' + B_\rho, K' \subseteq K + B_\rho\},$$

where  $B_\rho = \{x \in \mathbb{R}^n : \|x\| \leq \rho\}$ . A sequence  $K_j$  of compact subsets of  $\mathbb{R}^n$  converges to a compact subset  $K$  of  $\mathbb{R}^n$  if  $h(K_j, K) \rightarrow 0$  as  $j \rightarrow \infty$ .

It was shown by Hausdorff (1927) that any uniformly bounded sequence of compact subsets of  $\mathbb{R}^n$  has a convergent subsequence. In particular, any uniformly bounded sequence of compact convex subsets of  $\mathbb{R}^n$  has a subsequence which converges to a compact convex set. This special case of Hausdorff's result, which is all that we will later require, had already been established by Blaschke (1916) and is known as *Blaschke's selection principle*.

**Proposition 20** *Let  $\{A_k\}$  be a sequence of lattices in  $\mathbb{R}^n$  and let  $V_k$  be the Voronoi cell of  $A_k$ . If there exists a compact convex set  $V$  with nonempty interior such that  $V_k \rightarrow V$  in the Hausdorff metric as  $k \rightarrow \infty$ , then  $V$  is the Voronoi cell of a lattice  $A$  and  $A_k \rightarrow A$  as  $k \rightarrow \infty$ .*

*Proof* Since every Voronoi cell  $V_k$  is symmetric, so also is the limit  $V$ . Since  $V$  has nonempty interior, it follows that the origin is itself an interior point of  $V$ . Thus there exists  $\delta > 0$  such that the ball  $B_\delta = \{x \in \mathbb{R}^n : \|x\| \leq \delta\}$  is contained in  $V$ .

It follows that  $B_{\delta/2} \subseteq V_k$  for all large  $k$ . The quickest way to see this is to use *Rådström's cancellation law*, which says that if  $A, B, C$  are nonempty compact convex subsets of  $\mathbb{R}^n$  such that  $A + C \subseteq B + C$ , then  $A \subseteq B$ . In the present case we have

$$B_{\delta/2} + B_{\delta/2} \subseteq B_\delta \subseteq V \subseteq V_k + B_{\delta/2} \text{ for } k \geq k_0,$$

and hence  $B_{\delta/2} \subseteq V_k$  for  $k \geq k_0$ . Since also  $V_k \subseteq V + B_{\delta/2}$  for all large  $k$ , there exists  $R > 0$  such that  $V_k \subseteq B_R$  for all  $k$ .

The lattice  $\mathcal{A}_k$  has at most  $2(2^n - 1)$  facet vectors, by Proposition 15. Hence, by restriction to a subsequence, we may assume that all  $\mathcal{A}_k$  have the same number  $m$  of facet vectors. Let  $x_{k1}, \dots, x_{km}$  be the facet vectors of  $\mathcal{A}_k$  and choose the notation so that  $x_{k1}, \dots, x_{kn}$  are linearly independent. Since they all lie in the ball  $B_{2R}$ , by restriction to a further subsequence we may assume that

$$x_{kj} \rightarrow x_j \quad \text{as } k \rightarrow \infty \quad (j = 1, \dots, m).$$

Evidently  $\|x_j\| \geq \delta$  ( $j = 1, \dots, m$ ) since, for  $k \geq k_0$ , all nonzero  $x \in \mathcal{A}_k$  have  $\|x\| \geq \delta$ .

The set  $\mathcal{A}$  of all integral linear combinations of  $x_1, \dots, x_m$  is certainly an additive subgroup of  $\mathbb{R}^n$ . Moreover  $\mathcal{A}$  is discrete. For suppose  $y \in \mathcal{A}$  and  $\|y\| < \delta$ . We have

$$y = \alpha_1 x_1 + \dots + \alpha_m x_m,$$

where  $\alpha_j \in \mathbb{Z}$  ( $j = 1, \dots, m$ ). If

$$y_k = \alpha_1 x_{k1} + \dots + \alpha_m x_{km},$$

then  $y_k \rightarrow y$  as  $k \rightarrow \infty$  and hence  $\|y_k\| < \delta$  for all large  $k$ . Since  $y_k \in \mathcal{A}_k$ , it follows that  $y_k = O$  for all large  $k$  and hence  $y = O$ .

Since the lattice  $\mathcal{A}'_k$  with basis  $x_{k1}, \dots, x_{kn}$  is a sublattice of  $\mathcal{A}_k$ , we have

$$d(\mathcal{A}'_k) \geq d(\mathcal{A}_k) = \lambda(V_k) \geq \lambda(B_{\delta/2}).$$

Since  $d(\mathcal{A}'_k) = |\det(x_{k1}, \dots, x_{kn})|$ , it follows that also

$$|\det(x_1, \dots, x_n)| \geq \lambda(B_{\delta/2}) > 0.$$

Thus the vectors  $x_1, \dots, x_n$  are linearly independent. Hence  $\mathcal{A}$  is a lattice.

Let  $b_1, \dots, b_n$  be a basis of  $\mathcal{A}$ . Then, by the definition of  $\mathcal{A}$ ,

$$b_i = \alpha_{i1} x_1 + \dots + \alpha_{im} x_m,$$

where  $\alpha_{ij} \in \mathbb{Z}$  ( $1 \leq i \leq n, 1 \leq j \leq m$ ). Put

$$b_{ki} = \alpha_{i1} x_{k1} + \dots + \alpha_{im} x_{km}.$$

Then  $b_{ki} \in \mathcal{A}_k$  and  $b_{ki} \rightarrow b_i$  as  $k \rightarrow \infty$  ( $i = 1, \dots, n$ ). Hence, for all large  $k$ , the vectors  $b_{k1}, \dots, b_{kn}$  are linearly independent. We are going to show that  $b_{k1}, \dots, b_{kn}$  is a basis of  $\mathcal{A}_k$  for all large  $k$ .

Since  $b_1, \dots, b_n$  is a basis of  $\mathcal{A}$ , we have

$$x_j = \gamma_{j1} b_1 + \dots + \gamma_{jn} b_n,$$

where  $\gamma_{ji} \in \mathbb{Z}$  ( $1 \leq i \leq n, 1 \leq j \leq m$ ). Hence, if

$$y_{kj} = \gamma_{j1} b_{k1} + \dots + \gamma_{jn} b_{kn},$$

then  $y_{kj} \in \mathcal{A}_k$  and  $y_{kj} \rightarrow x_j$  as  $k \rightarrow \infty$  ( $j = 1, \dots, m$ ). Thus, for all large  $k$ ,

$$\|y_{kj} - x_{kj}\| < \delta \quad (j = 1, \dots, m).$$

Since  $y_{kj} - x_{kj} \in A_k$ , this implies that, for all large  $k$ ,  $y_{kj} = x_{kj}$  ( $j = 1, \dots, m$ ). Thus every facet vector of  $A_k$  is an integral linear combination of  $b_{k1}, \dots, b_{kn}$  and hence, by Proposition 16, every vector of  $A_k$  is an integral linear combination of  $b_{k1}, \dots, b_{kn}$ . Since  $b_{k1}, \dots, b_{kn}$  are linearly independent, this shows that they are a basis of  $A_k$ .

Let  $W$  be the Voronoi cell of  $A$ . We wish to show that  $V = W$ . If  $v \in V$ , then there exist  $v_k \in V_k$  such that  $v_k \rightarrow v$ . Assume  $v \notin W$ . Then  $\|v\| > \|z - v\|$  for some  $z \in A$ , and so

$$\|v\| = \|z - v\| + \rho,$$

where  $\rho > 0$ . There exist  $z_k \in A_k$  such that  $z_k \rightarrow z$ . Then, for all large  $k$ ,

$$\|v\| > \|z_k - v\| + \rho/2$$

and hence, for all large  $k$ ,

$$\|v_k\| > \|z_k - v_k\|.$$

But this contradicts  $v_k \in V_k$ .

This proves that  $V \subseteq W$ . On the other hand,  $V$  has volume

$$\begin{aligned} \lambda(V) &= \lim_{k \rightarrow \infty} \lambda(V_k) = \lim_{k \rightarrow \infty} d(A_k) \\ &= \lim_{k \rightarrow \infty} |\det(b_{k1}, \dots, b_{kn})| \\ &= |\det(b_1, \dots, b_n)| = d(A) = \lambda(W). \end{aligned}$$

It follows that every interior point of  $W$  is in  $V$ , and hence  $W = V$ . Corollary 17 now shows that the same lattice  $A$  would have been obtained if we had restricted attention to some other subsequence of  $\{A_k\}$ .

Let  $a_1, \dots, a_n$  be any basis of  $A$ . We are going to show that, for the sequence  $\{A_k\}$  originally given, there exist  $a_{ki} \in A_k$  such that

$$a_{ki} \rightarrow a_i \text{ as } k \rightarrow \infty \quad (i = 1, \dots, n).$$

If this is not the case then, for some  $i \in \{1, \dots, n\}$  and some  $\varepsilon > 0$ , there exist infinitely many  $k$  such that

$$\|x - a_i\| > \varepsilon \quad \text{for all } x \in A_k.$$

From this subsequence we could as before pick a further subsequence  $A_{k_v} \rightarrow A$ . Then every  $y \in A$  is the limit of a sequence  $y_v \in A_{k_v}$ . Taking  $y = a_i$ , we obtain a contradiction.

It only remains to show that  $a_{k1}, \dots, a_{kn}$  is a basis of  $A_k$  for all large  $k$ . Since

$$\begin{aligned} \lim_{k \rightarrow \infty} |\det(a_{k1}, \dots, a_{kn})| &= |\det(a_1, \dots, a_n)| \\ &= d(A) = \lambda(V) = \lim_{k \rightarrow \infty} \lambda(V_k), \end{aligned}$$

for all large  $k$  we must have

$$0 < |\det(a_{k1}, \dots, a_{kn})| < 2\lambda(V_k).$$

But if  $a_{k1}, \dots, a_{kn}$  were not a basis of  $A_k$  for all large  $k$ , then for infinitely many  $k$  we would have

$$|\det(a_{k1}, \dots, a_{kn})| \geq 2d(A_k) = 2\lambda(V_k). \quad \square$$

Proposition 20 has the following counterpart:

**Proposition 21** *Let  $\{A_k\}$  be a sequence of lattices in  $\mathbb{R}^n$  and let  $V_k$  be the Voronoi cell of  $A_k$ . If there exists a lattice  $A$  such that  $A_k \rightarrow A$  as  $k \rightarrow \infty$ , and if  $V$  is the Voronoi cell of  $A$ , then  $V_k \rightarrow V$  in the Hausdorff metric as  $k \rightarrow \infty$ .*

*Proof* By hypothesis, there exists a basis  $b_1, \dots, b_n$  of  $A$  and a basis  $b_{k1}, \dots, b_{kn}$  of each  $A_k$  such that  $b_{kj} \rightarrow b_j$  as  $k \rightarrow \infty$  ( $j = 1, \dots, n$ ). Choose  $R > 0$  so that the fundamental parallelotope of  $A$  is contained in the ball  $B_R = \{x \in \mathbb{R}^n : \|x\| \leq R\}$ . Then, for all  $k \geq k_0$ , the fundamental parallelotope of  $A_k$  is contained in the ball  $B_{2R}$ . It follows that, for all  $k \geq k_0$ , every point of  $\mathbb{R}^n$  is distant at most  $2R$  from some point of  $A_k$  and hence  $V_k \subseteq B_{2R}$ .

Consequently, by Blaschke's selection principle, the sequence  $\{V_k\}$  has a subsequence  $\{V_{k_v}\}$  which converges in the Hausdorff metric to a compact convex set  $W$ . Moreover,

$$\lambda(W) = \lim_{v \rightarrow \infty} \lambda(V_{k_v}) = \lim_{v \rightarrow \infty} d(A_{k_v}) = d(A) > 0.$$

Consequently, since  $W$  is convex, it has nonempty interior. It now follows from Proposition 20 that  $W = V$ .

Thus any convergent subsequence of  $\{V_k\}$  has the same limit  $V$ . If the whole sequence  $\{V_k\}$  did not converge to  $V$ , there would exist  $\rho > 0$  and a subsequence  $\{V_{k_v}\}$  such that

$$h(V_{k_v}, V) \geq \rho \quad \text{for all } v.$$

By the Blaschke selection principle again, this subsequence would itself have a convergent subsequence. Since its limit must be  $V$ , this yields a contradiction.  $\square$

Suppose  $A_k \in \mathcal{L}_n$  and  $A_k \rightarrow A$  as  $k \rightarrow \infty$ . We will show that not only  $d(A_k) \rightarrow d(A)$ , but also  $m(A_k) \rightarrow m(A)$  as  $k \rightarrow \infty$ . Since every  $x \in A$  is the limit of a sequence  $x_k \in A_k$ , we must have  $\varliminf_{k \rightarrow \infty} m(A_k) \leq m(A)$ . On the other hand, by Proposition 19, if  $x_k \in A_k$  and  $x_k \rightarrow x$ , then  $x \in A$ . Hence  $\varliminf_{k \rightarrow \infty} m(A) \geq m(A)$ , since  $x \neq 0$  if  $x_k \neq 0$  for large  $k$ .

Suppose now that a subset  $\mathcal{F}$  of  $\mathcal{L}_n$  has the property that any infinite sequence  $A_k$  of lattices in  $\mathcal{F}$  has a convergent subsequence. Then there exist positive constants  $\rho, \sigma$  such that

$$m(A) \geq \rho^2, \quad d(A) \leq \sigma \quad \text{for all } A \in \mathcal{F}.$$

For otherwise there would exist a sequence  $\Lambda_k$  of lattices in  $\mathcal{F}$  such that either  $m(\Lambda_k) \rightarrow 0$  or  $d(\Lambda_k) \rightarrow \infty$ , and clearly this sequence could have no convergent subsequence.

We now prove the fundamental *compactness theorem* of Mahler (1946), which says that this necessary condition on  $\mathcal{F}$  is also sufficient.

**Proposition 22** *If  $\{\Lambda_k\}$  is a sequence of lattices in  $\mathbb{R}^n$  such that*

$$m(\Lambda_k) \geq \rho^2, \quad d(\Lambda_k) \leq \sigma \quad \text{for all } k,$$

*where  $\rho, \sigma$  are positive constants, then the sequence  $\{\Lambda_k\}$  certainly has a convergent subsequence.*

*Proof* Let  $V_k$  denote the Voronoi cell of  $\Lambda_k$ . We show first that the ball  $B_{\rho/2} = \{x \in \mathbb{R}^n : \|x\| \leq \rho/2\}$  is contained in every Voronoi cell  $V_k$ . In fact if  $\|x\| \leq \rho/2$  then, for every nonzero  $y \in \Lambda_k$ ,

$$\|x - y\| \geq \|y\| - \|x\| \geq \rho - \rho/2 = \rho/2 \geq \|x\|,$$

and hence  $x \in V_k$ .

Let  $v_k$  be a point of  $V_k$  which is furthest from the origin. Then  $V_k$  contains the convex hull  $C_k$  of the set  $v_k \cup B_{\rho/2}$ . Since the volume of  $V_k$  is bounded above by  $\sigma$ , so also is the volume of  $C_k$ . But this implies that the sequence  $v_k$  is bounded. Thus there exists  $R > 0$  such that the ball  $B_R$  contains every Voronoi cell  $V_k$ .

By Blaschke's selection principle, the sequence  $\{V_k\}$  has a subsequence  $\{V_{k_v}\}$  which converges in the Hausdorff metric to a compact convex set  $V$ . Since  $B_{\rho/2} \subseteq V$ , it follows from Proposition 20 that  $\Lambda_{k_v} \rightarrow \Lambda$ , where  $\Lambda$  is a lattice with Voronoi cell  $V$ .  $\square$

To illustrate the utility of Mahler's compactness theorem, we now show that, as stated in Section 3, any compact symmetric convex set  $K$  with nonempty interior has a critical lattice.

By the definition of the critical determinant  $\Delta(K)$ , there exists a sequence  $\Lambda_k$  of lattices with no nonzero points in the interior of  $K$  such that  $d(\Lambda_k) \rightarrow \Delta(K)$  as  $k \rightarrow \infty$ . Since  $K$  contains a ball  $B_\rho$  with radius  $\rho > 0$ , we have  $m(\Lambda_k) \geq \rho^2$  for all  $k$ . Hence, by Proposition 22, there is a subsequence  $\Lambda_{k_v}$  which converges to a lattice  $\Lambda$  as  $v \rightarrow \infty$ . Since every point of  $\Lambda$  is a limit of points of  $\Lambda_{k_v}$ , no nonzero point of  $\Lambda$  lies in the interior of  $K$ . Furthermore,

$$d(\Lambda) = \lim_{v \rightarrow \infty} d(\Lambda_{k_v}) = \Delta(K),$$

and hence  $\Lambda$  is a critical lattice for  $K$ .

## 7 Further Remarks

The geometry of numbers is treated more extensively in Cassels [11], Erdős *et al.* [22] and Gruber and Lekkerkerker [27]. Minkowski's own account is available in [42].

Numerous references to the earlier literature are given in Keller [34]. Lagarias [36] gives an overview of lattice theory. For a simple proof that the indicator function of a convex set is Riemann integrable, see Szabo [57].

Diophantine approximation is studied in Cassels [12], Koksma [35] and Schmidt [50]. Minkowski's result that the discriminant of an algebraic number field other than  $\mathbb{Q}$  has absolute value greater than 1 is proved in Narkiewicz [44], for example.

Minkowski's theorem on successive minima is proved in Bambah *et al.* [3]. For the results of Banaszczyk mentioned in §3, see [4] and [5]. Sharp forms of Siegel's lemma are proved not only in Bombieri and Vaaler [7], but also in Matveev [40]. The result of Gillet and Soulé appeared in [25]. Some interesting results and conjectures concerning the product  $\lambda(K)\lambda(K^*)$  are described on pp. 425–427 of Schneider [51].

An algorithm of Lovász, which first appeared in Lenstra, Lenstra and Lovász [38], produces in finitely many steps a basis for a lattice  $\Lambda$  in  $\mathbb{R}^n$  which is 'reduced'. Although the first vector of a reduced basis is in general not a minimal vector, it has square-norm at most  $2^{n-1}m(\Lambda)$ . This suffices for many applications and the algorithm has been used to solve a number of apparently unrelated computational problems, such as factoring polynomials in  $\mathbb{Q}[t]$ , integer linear programming and simultaneous Diophantine approximation. There is an account of the basis reduction algorithm in Schrijver [52]. The algorithmic geometry of numbers is surveyed in Kannan [33].

Mahler [39] has established an analogue of the geometry of numbers for formal Laurent series with coefficients from an arbitrary field  $F$ , the roles of  $\mathbb{Z}$ ,  $\mathbb{Q}$  and  $\mathbb{R}$  being taken by  $F[t]$ ,  $F(t)$  and  $F((t))$ . In particular, Eichler [19] has shown that the Riemann–Roch theorem for algebraic functions may be thus derived by geometry of numbers arguments.

There is also a generalization of Minkowski's lattice point theorem to locally compact groups, with Haar measure taking the place of volume; see Chapter 2 (Lemma 1) of Weil [60].

Voronoi *diagrams* and their uses are surveyed in Aurenhammer [1]. Proofs of the basic properties of polytopes referred to in §4 may be found in Brøndsted [9] and Coppel [15]. Planar tilings are studied in detail in Grünbaum and Shephard [28].

Mathematical crystallography is treated in Schwarzenberger [53] and Engel [21]. For the physicist's point of view, see Burckhardt [10], Janssen [32] and Birman [6]. There is much theoretical information, in addition to tables, in [31].

For Bieberbach's theorems, see Vince [59], Charlap [13] and Milnor [41]. Various equivalent forms for the definitions of crystal and crystallographic group are given in Dolbilin *et al.* [17]. It is shown in Charlap [13] that crystallographic groups may be abstractly characterized as groups containing a finitely generated maximal abelian torsion-free subgroup of finite index. (An abelian group is *torsion-free* if only the identity element has finite order.) The fundamental group of a compact flat Riemannian manifold is a torsion-free crystallographic group and all torsion-free crystallographic groups may be obtained in this way. For these connections with differential geometry, see Wolf [61] and Charlap [13].

In more than 4 dimensions the complete enumeration of all crystallographic groups is no longer practicable. However, algorithms for deciding if two crystallographic groups are equivalent in some sense have been developed by Opgenorth *et al.* [45].

An interesting subset of all crystallographic groups consists of those generated by reflections in hyperplanes, since Stiefel (1941/2) showed that they are in 1-1 correspondence with the compact simply-connected semi-simple Lie groups. See the ‘Note historique’ in Bourbaki [8].

There has recently been considerable interest in tilings of  $\mathbb{R}^n$  which, although not lattice tilings, consist of translates of finitely many  $n$ -dimensional polytopes. The first example, in  $\mathbb{R}^2$ , due to Penrose (1974), was explained more algebraically by de Bruijn (1981). A substantial generalization of de Bruijn’s construction was given by Katz and Duneau (1986), who showed that many such ‘quasiperiodic’ tilings may be obtained by a method of cut and projection from ordinary lattices in a higher-dimensional space. The subject gained practical significance with the discovery by Shechtman *et al.* (1984) that the diffraction pattern of an alloy of aluminium and magnesium has icosahedral symmetry, which is impossible for a crystal. Many other ‘quasicrystals’ have since been found. The papers referred to are reproduced, with others, in Steinhardt and Ostlund [56]. The mathematical theory of quasicrystals is surveyed in Le *et al.* [37].

Skubenko [54] has given an upper bound for Hermite’s constant  $\gamma_n$ . Somewhat sharper bounds are known, but they have the same asymptotic behaviour and the proofs are much more complicated. A lower bound for  $\gamma_n$  was obtained with a new method by Ball [2].

For the densest lattices in  $\mathbb{R}^n$  ( $n \leq 8$ ), see Rysikov and Baranovskii [49]. The enumeration of all root lattices is carried out in Ebeling [18]. (A more general problem is treated in Chap. 3 of Humphreys [30] and in Chap. 6 of Bourbaki [8].) For the Voronoi cells of root lattices, see Chap. 21 of Conway and Sloane [14] and Moody and Patera [43]. For the *Dynkin diagrams* associated with root lattices, see also Reiten [47].

Rajan and Shende [46] characterize root lattices as those lattices for which every facet vector is a minimal vector, but their definition of root lattice is not that adopted here. Their argument shows that if every facet vector of a lattice is a minimal vector then, after scaling to make the minimal vectors have square-norm 2, it is a root lattice in our sense.

There is a fund of information about lattice packings of balls in Conway and Sloane [14]. See also Thompson [58] for the Leech lattice and Coxeter [16] for the kissing number problem.

We have restricted attention to lattice packings and, in particular, to lattice packings of balls. Lattice packings of other convex bodies are discussed in the books on geometry of numbers cited above. Non-lattice packings have also been much studied. The notion of density is not so intuitive in this case and it should be realized that the density is unaltered if finitely many sets are removed from the packing.

Packings and coverings are discussed in the texts of Rogers [48] and Fejes Tóth [23], [24]. For packings of balls, see also Zong [62]. Sloane [55] and Elkies [20] provide introductions to the connections between lattice packings of balls and coding theory.

The third part of Hilbert’s 18th problem, which is surveyed in Milnor [41], deals with the densest lattice or non-lattice packing of balls in  $\mathbb{R}^n$ . It is known that, for  $n = 2$ , the densest lattice packing is also a densest packing. The original proof by Thue (1882/1910) was incomplete, but a complete proof was given by L. Fejes Tóth (1940). The famous *Kepler conjecture* asserts that, also for  $n = 3$ , the densest lattice

packing is a densest packing. A computer-aided proof has recently been announced by Hales [29]. It is unknown if the same holds for any  $n > 3$ .

Propositions 20 and 21 are due to Groemer [26], and are of interest quite apart from the application to Mahler's compactness theorem. Other proofs of the latter are given in Cassels [11] and Gruber and Lekkerkerker [27]. Blaschke's selection principle and Rådström's cancellation law are proved in [15] and [51], for example.

## 8 Selected References

- [1] F. Aurenhammer, Voronoi diagrams – a survey of a fundamental geometric data structure, *ACM Computing Surveys* **23** (1991), 345–405.
- [2] K. Ball, A lower bound for the optimal density of lattice packings, *Internat. Math. Res. Notices* 1992, no. 10, 217–221.
- [3] R.P. Bambah, A.C. Woods and H. Zassenhaus, Three proofs of Minkowski's second inequality in the geometry of numbers, *J. Austral. Math. Soc.* **5** (1965), 453–462.
- [4] W. Banaszczyk, New bounds in some transference theorems in the geometry of numbers, *Math. Ann.* **296** (1993), 625–635.
- [5] W. Banaszczyk, Inequalities for convex bodies and polar reciprocal lattices in  $\mathbb{R}^n$ . II. Application of  $K$ -convexity, *Discrete Comput. Geom.* **16** (1996), 305–311.
- [6] J.L. Birman, *Theory of crystal space groups and lattice dynamics*, Springer-Verlag, Berlin, 1984. [Original edition in *Handbuch der Physik*, 1974]
- [7] E. Bombieri and J. Vaaler, On Siegel's lemma, *Invent. Math.* **73** (1983), 11–32.
- [8] N. Bourbaki, *Groupes et algèbres de Lie, Chapitres 4,5 et 6*, Masson, Paris, 1981.
- [9] A. Brøndsted, *An introduction to convex polytopes*, Springer-Verlag, New York, 1983.
- [10] J.J. Burckhardt, *Die Bewegungsgruppen der Kristallographie*, 2nd ed., Birkhäuser, Basel, 1966.
- [11] J.W.S. Cassels, *An introduction to the geometry of numbers*, corrected reprint, Springer-Verlag, Berlin, 1997. [Original edition, 1959]
- [12] J.W.S. Cassels, *An introduction to Diophantine approximation*, Cambridge University Press, 1957.
- [13] L.S. Charlap, *Bieberbach groups and flat manifolds*, Springer-Verlag, New York, 1986.
- [14] J.H. Conway and N.J.A. Sloane, *Sphere packings, lattices and groups*, 3rd ed., Springer-Verlag, New York, 1999.
- [15] W.A. Coppel, *Foundations of convex geometry*, Cambridge University Press, 1998.
- [16] H.S.M. Coxeter, An upper bound for the number of equal nonoverlapping spheres that can touch another of the same size, *Convexity* (ed. V. Klee), pp. 53–71, Proc. Symp. Pure Math. **7**, Amer. Math. Soc., Providence, Rhode Island, 1963.
- [17] N.P. Dolbilin, J.C. Lagarias and M. Senechal, Multiregular point systems, *Discrete Comput. Geom.* **20** (1998), 477–498.
- [18] W. Ebeling, *Lattices and codes*, Vieweg, Braunschweig, 1994.
- [19] M. Eichler, Ein Satz über Linearformen in Polynombereichen, *Arch. Math.* **10** (1959), 81–84.
- [20] N.D. Elkies, Lattices, linear codes, and invariants, *Notices Amer. Math. Soc.* **47** (2000), 1238–1245 and 1382–1391.
- [21] P. Engel, Geometric crystallography, *Handbook of convex geometry* (ed. P.M. Gruber and J.M. Wills), Volume B, pp. 989–1041, North-Holland, Amsterdam, 1993. (The same volume contains several other useful survey articles relevant to this chapter.)
- [22] P. Erdős, P.M. Gruber and J. Hammer, *Lattice points*, Longman, Harlow, Essex, 1989.
- [23] L. Fejes Tóth, *Regular Figures*, Pergamon, Oxford, 1964.