[24] L. Fejes Tóth, *Lagerungen in der Ebene auf der Kugel und im Raum*, 2nd ed., Springer-Verlag, Berlin, 1972.
[25] H. Gillet and C. Soulé, On the number of lattice points in convex symmetric bodies and their duals, *Israel J. Math.* **74** (1991), 347–357.
[26] H. Groemer, Continuity properties of Voronoi domains, *Monatsh. Math.* **75** (1971), 423–431.
[27] P.M. Gruber and C.G. Lekkerkerker, *Geometry of numbers*, 2nd ed., North-Holland, Amsterdam, 1987.
[28] B. Grünbaum and G.C. Shephard, *Tilings and patterns*, Freeman, New York, 1987.
[29] T.C. Hales, Cannonballs and honeycombs, *Notices Amer. Math. Soc.* **47** (2000), 440–449.
[30] J.E. Humphreys, *Introduction to Lie algebras and representation theory*, Springer-Verlag, New York, 1972.
[31] *International tables for crystallography, Vols. A-C*, Kluwer, Dordrecht, 1983–1993.
[32] T. Janssen, *Crystallographic groups*, North-Holland, Amsterdam, 1973.
[33] R. Kannan, Algorithmic geometry of numbers, *Annual review of computer science* **2** (1987), 231–267.
[34] O.-H. Keller, *Geometrie der Zahlen*, Enzyklopädie der mathematischen Wissenschaften I-2, 27, Teubner, Leipzig, 1954.
[35] J.F. Koksma, *Diophantische Approximationen*, Springer-Verlag, Berlin, 1936. [Reprinted Chelsea, New York, 1950]
[36] J.C. Lagarias, Point lattices, *Handbook of Combinatorics* (ed. R. Graham, M. Grötschel and L. Lovász), Vol. I, pp. 919–966, Elsevier, Amsterdam, 1995.
[37] T.Q.T. Le, S.A. Piunikhin and V.A. Sadov, The geometry of quasicrystals, *Russian Math. Surveys* **48** (1993), no. 1, 37–100.
[38] A.K. Lenstra, H.W. Lenstra and L. Lovász, Factoring polynomials with rational coefficients, *Math. Ann.* **261** (1982), 515–534.
[39] K. Mahler, An analogue to Minkowski's geometry of numbers in a field of series, *Ann. of Math.* **42** (1941), 488–522.
[40] E.M. Matveev, On linear and multiplicative relations, *Math. USSR-Sb.* **78** (1994), 411–425.
[41] J. Milnor, Hilbert's Problem 18: On crystallographic groups, fundamental domains, and on sphere packing, *Mathematical developments arising from Hilbert problems* (ed. F.E. Browder), pp. 491–506, Proc. Symp. Pure Math. **28**, Part 2, Amer. Math. Soc., Providence, Rhode Island, 1976.
[42] H. Minkowski, *Geometrie der Zahlen*, Teubner, Leipzig, 1896. [Reprinted Chelsea, New York, 1953]
[43] R.V. Moody and J. Patera, Voronoi and Delaunay cells of root lattices: classification of their faces and facets by Coxeter–Dynkin diagrams, *J. Phys. A* **25** (1992), 5089–5134.
[44] W. Narkiewicz, *Elementary and analytic theory of algebraic numbers*, 2nd ed., Springer-Verlag, Berlin, 1990.
[45] J. Opgenorth, W. Plesken and T. Schulz, Crystallographic algorithms and tables, *Acta Cryst. A* **54** (1998), 517–531.
[46] D.S. Rajan and A.M. Shende, A characterization of root lattices, *Discrete Math.* **161** (1996), 309–314.
[47] I. Reiten, Dynkin diagrams and the representation theory of Lie algebras, *Notices Amer. Math. Soc.* **44** (1997), 546–556.
[48] C.A. Rogers, *Packing and covering*, Cambridge University Press, 1964.
[49] S.S. Ryshkov and E.P. Baranovskii, Classical methods in the theory of lattice packings, *Russian Math. Surveys* **34** (1979), no. 4, 1–68.
[50] W.M. Schmidt, *Diophantine approximation*, Lecture Notes in Mathematics **785**, Springer-Verlag, Berlin, 1980.
[51] R. Schneider, *Convex bodies*: *the Brunn–Minkowski theory*, Cambridge University Press, 1993.

[52] A. Schrijver, *Theory of linear and integer programming*, corrected reprint, Wiley, Chichester, 1989.
[53] R.L.E. Schwarzenberger, *N-dimensional crystallography*, Pitman, London, 1980.
[54] B.F. Skubenko, A remark on an upper bound on the Hermite constant for the densest lattice packings of spheres, *J. Soviet Math.* **18** (1982), 960–961.
[55] N.J.A. Sloane, The packing of spheres, *Scientific American* **250** (1984), 92–101.
[56] P.J. Steinhardt and S. Ostlund (ed.), *The physics of quasicrystals*, World Scientific, Singapore, 1987.
[57] L. Szabo, A simple proof for the Jordan measurability of convex sets, *Elem. Math.* **52** (1997), 84–86.
[58] T.M. Thompson, *From error-correcting codes through sphere packings to simple groups*, Carus Mathematical Monograph No. 21, Mathematical Association of America, 1983.
[59] A. Vince, Periodicity, quasiperiodicity and Bieberbach's theorem, *Amer. Math. Monthly* **104** (1997), 27–35.
[60] A. Weil, *Basic number theory*, 2nd ed., Springer-Verlag, Berlin, 1973.
[61] J.A. Wolf, *Spaces of constant curvature*, 3rd ed., Publish or Perish, Boston, Mass., 1974.
[62] C. Zong, *Sphere packings*, Springer-Verlag, New York, 1999.

Additional References

F. Pfender and G. Ziegler, Kissing numbers, sphere packings and some unexpected proofs, *Notices Amer. Math. Soc.* **51** (2004), 873–883. [The Leech lattice is indeed the densest lattice in $\mathbb{R}^{24}$.]

O.R. Musin, The problem of the twenty-five spheres, *Russian Math. Surveys* **58** (2003), 794–795. [The kissing number of $\mathbb{R}^4$ is 24.]

G. Muraz and J.-L. Verger-Gaugry, On a generalization of the selection theorem of Mahler, *Journal de Théorie des Nombres de Bordeaux* **17** (2005), 237–269. [Extends Mahler's compactness theorem for lattices to sets which are uniformly discrete and uniformly relatively dense.]

IX

The Number of Prime Numbers

1 Finding the Problem

It was already shown in Euclid's *Elements* (Book IX, Proposition 20) that there are infinitely many prime numbers. The proof is a model of simplicity: let $p_1, \ldots, p_n$ be any finite set of primes and consider the integer $N = p_1 \cdots p_n + 1$. Then $N > 1$ and each prime divisor p of N is distinct from $p_1, \ldots, p_n$, since $p = p_j$ would imply that p divides $N - p_1 \cdots p_n = 1$. It is worth noting that the same argument applies if we take $N = p_1^{\alpha_1} \cdots p_n^{\alpha_n} + 1$, with any positive integers $\alpha_1, \ldots, \alpha_n$.

Euler (1737) gave an analytic proof of Euclid's result, which provides also quantitative information about the distribution of primes:

Proposition 1 *The series $\sum_p 1/p$, where p runs through all primes, is divergent.*

Proof For any prime p we have

$$(1 - 1/p)^{-1} = 1 + p^{-1} + p^{-2} + \cdots$$

and hence

$$\prod_{p \le x} (1 - 1/p)^{-1} = \prod_{p \le x} (1 + p^{-1} + p^{-2} + \cdots) > \sum_{n \le x} 1/n,$$

since any positive integer $n \le x$ is a product of powers of primes $p \le x$. Since

$$\sum_{n \le x} 1/n > \sum_{n \le x} \int_n^{n+1} dt/t > \log x,$$

it follows that

$$\prod_{p \le x} (1 - 1/p)^{-1} > \log x.$$

On the other hand, since the representation of any positive integer as a product of prime powers is *unique*,

$$\prod_{p \le x} (1 - 1/p^2)^{-1} = \prod_{p \le x} (1 + p^{-2} + p^{-4} + \cdots) \le \sum_{n=1}^{\infty} 1/n^2 =: S,$$

W.A. Coppel, *Number Theory: An Introduction to Mathematics*, Universitext,
DOI: 10.1007/978-0-387-89486-7_9, © Springer Science + Business Media, LLC 2009

and

$$S = 1 + \sum_{n=1}^{\infty} 1/(n+1)^2 < 1 + \sum_{n=1}^{\infty} \int_n^{n+1} dt/t^2 = 1 + \int_1^{\infty} dt/t^2 = 2.$$

(In fact $S = \pi^2/6$, as Euler (1735) also showed.) Since $1-1/p^2 = (1-1/p)(1+1/p)$, and since $1 + x \leq e^x$, it follows that

$$\prod_{p \leq x} (1 - 1/p)^{-1} \leq S \prod_{p \leq x} (1 + 1/p) < S\, e^{\sum_{p \leq x} 1/p}.$$

Combining this with the inequality of the previous paragraph, we obtain

$$\sum_{p \leq x} 1/p > \log\log x - \log S. \qquad \square$$

Since the series $\sum_{n=1}^{\infty} 1/n^2$ is convergent, Proposition 1 says that 'there are more primes than squares'. Proposition 1 can be made more precise. It was shown by Mertens (1874) that

$$\sum_{p \leq x} 1/p = \log\log x + c + O(1/\log x),$$

where c is a constant ($c = 0.261497\ldots$).

Let $\pi(x)$ denote the number of primes $\leq x$:

$$\pi(x) = \sum_{p \leq x} 1.$$

It may be asked whether $\pi(x)$ has some simple asymptotic behaviour as $x \to \infty$. It is not obvious that this is a sensible question. The behaviour of $\pi(x)$ for small values of x is quite irregular. Moreover the sequence of positive integers contains arbitrarily large blocks without primes; for example, none of the integers

$$n! + 2, n! + 3, \ldots, n! + n$$

is a prime. Indeed Euler (1751) expressed the view that "there reigns neither order nor rule" in the sequence of prime numbers.

From an analysis of tables of primes Legendre (1798) was led to conjecture that, for large values of x, $\pi(x)$ is given approximately by the formula

$$x/(A \log x - B),$$

where A, B are constants and $\log x$ again denotes the natural logarithm of x (i.e., to the base e). In 1808 he proposed the specific values $A = 1$, $B = 1.08366$.

The first significant results on the asymptotic behaviour of $\pi(x)$ were obtained by Chebyshev (1849). He proved that, for each positive integer n,

$$\underline{\lim}_{x\to\infty} \left(\pi(x) - \int_2^x dt/\log t\right) \log^n x/x \leq 0$$

$$\leq \overline{\lim}_{x\to\infty} \left(\pi(x) - \int_2^x dt/\log t\right) \log^n x/x,$$

where $\log^n x = (\log x)^n$. By repeatedly integrating by parts it may be seen that, for each positive integer n,

$$\int_2^x dt/\log t = \{1 + 1!/\log x + 2!/\log^2 x + \cdots + (n-1)!/\log^{n-1} x\}x/\log x + n!\int_2^x dt/\log^{n+1} t + c_n,$$

where c_n is a constant. Moreover, using the *Landau order symbol* defined under 'Notations',

$$\int_2^x dt/\log^{n+1} t = O(x/\log^{n+1} x),$$

since

$$\int_2^{x^{1/2}} dt/\log^{n+1} t < x^{1/2}/\log^{n+1} 2, \qquad \int_{x^{1/2}}^x dt/\log^{n+1} t < 2^{n+1}x/\log^{n+1} x.$$

Thus Chebyshev's result shows that $A = B = 1$ are the best possible values for a formula of Legendre's type and suggests that

$$Li(x) = \int_2^x dt/\log t$$

is a better approximation to $\pi(x)$.

If we interpret this approximation as an asymptotic formula, then it implies that $\pi(x)\log x/x \to 1$ as $x \to \infty$, i.e., using another *Landau order symbol*,

$$\pi(x) \sim x/\log x. \tag{1}$$

The validity of the relation (1) is now known as the *prime number theorem*. If the n-th prime is denoted by p_n, then the prime number theorem can also be stated in the form $p_n \sim n\log n$:

Proposition 2 *$\pi(x) \sim x/\log x$ if and only if $p_n \sim n\log n$.*

Proof If $\pi(x)\log x/x \to 1$, then

$$\log\pi(x) + \log\log x - \log x \to 0$$

and hence

$$\log\pi(x)/\log x \to 1.$$

Consequently

$$\pi(x)\log\pi(x)/x = \pi(x)\log x/x \cdot \log\pi(x)/\log x \to 1.$$

Since $\pi(p_n) = n$, this shows that $p_n \sim n\log n$.

Conversely, suppose $p_n/n\log n \to 1$. Since

$$(n+1)\log(n+1)/n\log n = (1+1/n)\{1+\log(1+1/n)/\log n\} \to 1,$$

it follows that $p_{n+1}/p_n \to 1$. Furthermore

$$\log p_n - \log n - \log\log n \to 0,$$

and hence

$$\log p_n/\log n \to 1.$$

If $p_n \le x < p_{n+1}$, then $\pi(x) = n$ and

$$n\log p_n/p_{n+1} \le \pi(x)\log x/x \le n\log p_{n+1}/p_n.$$

Since

$$n\log p_n/p_{n+1} = p_n/p_{n+1}\cdot n\log n/p_n \cdot \log p_n/\log n \to 1$$

and similarly $n\log p_{n+1}/p_n \to 1$, it follows that also $\pi(x)\log x/x \to 1$. □

Numerical evidence, both for the prime number theorem and for the fact that $Li(x)$ is a better approximation than $x/\log x$ to $\pi(x)$, is provided by Table 1.

In a second paper Chebyshev (1852) made some progress towards proving the prime number theorem by showing that

$$a \le \underline{\lim_{x\to\infty}}\ \pi(x)\log x/x \le \overline{\lim_{x\to\infty}}\ \pi(x)\log x/x \le 6a/5,$$

where $a = 0.92129$. He used his results to give the first proof of *Bertrand's postulate*: for every real $x > 1$, there is a prime between x and $2x$.

New ideas were introduced by Riemann (1859), who linked the asymptotic behaviour of $\pi(x)$ with the behaviour of the function

$$\zeta(s) = \sum_{n=1}^{\infty} 1/n^s$$

Table 1.

x	$\pi(x)$	$x/\log x$	$Li(x)$	$\pi(x)\log x/x$	$\pi(x)/Li(x)$
10^3	168	144.	177.	1.16	0.94
10^4	1 229	1 085.	1 245.	1.132	0.987
10^5	9 592	8 685.	9 629.	1.1043	0.9961
10^6	78 498	72 382.	78 627.	1.08449	0.99835
10^7	664 579	620 420.	664 917.	1.07117	0.99949
10^8	5 761 455	5 428 681.	5 762 208.	1.06130	0.99987
10^9	50 847 534	48 254 942.	50 849 234.	1.05373	0.999966
10^{10}	455 052 511	434 294 481.	455 055 614.	1.04780	0.999993

for complex values of s. By developing these ideas, and by showing especially that $\zeta(s)$ has no zeros on the line $\mathscr{R}s = 1$, Hadamard and de la Vallée Poussin proved the prime number theorem (independently) in 1896. Shortly afterwards de la Vallée Poussin (1899) confirmed that $Li(x)$ was a better approximation than $x/\log x$ to $\pi(x)$ by proving (in particular) that

$$\pi(x) = Li(x) + O(x/\log^{\alpha} x) \quad \textit{for every } \alpha > 0. \tag{2}$$

Better error bounds than de la Vallée Poussin's have since been obtained, but they still fall far short of what is believed to be true.

Another approach to the prime number theorem was found by Wiener (1927–1933), as an application of his general theory of Tauberian theorems. A convenient form for this application was given by Ikehara (1931), and Bochner (1933) showed that in this case Wiener's general theory could be avoided.

It came as a great surprise to the mathematical community when in 1949 Selberg, assisted by Erdős, found a new proof of the prime number theorem which uses only the simplest facts of real analysis. Though elementary in a technical sense, this proof was still quite complicated. As a result of several subsequent simplifications it can now be given quite a clear and simple form. Nevertheless the Wiener–Ikehara proof will be presented here on account of its greater versatility. The error bound (2) can be obtained by both the Wiener and Selberg approaches, in the latter case at the cost of considerable complication.

2 Chebyshev's Functions

In his second paper Chebyshev introduced two functions

$$\theta(x) = \sum_{p \le x} \log p, \quad \psi(x) = \sum_{p^{\alpha} \le x} \log p,$$

which have since played a major role. Although $\psi(x)$ has the most complicated definition, it is easier to treat analytically than either $\theta(x)$ or $\pi(x)$. As we will show, the asymptotic behaviour of $\theta(x)$ is essentially the same as that of $\psi(x)$, and the asymptotic behaviour of $\pi(x)$ may be deduced without difficulty from that of $\theta(x)$.

Evidently

$$\theta(x) = \psi(x) = 0 \quad \text{for } x < 2$$

and

$$0 < \theta(x) \le \psi(x) \quad \text{for } x \ge 2.$$

Lemma 3 *The asymptotic behaviours of $\psi(x)$ and $\theta(x)$ are connected by*

(i) $\psi(x) - \theta(x) = O(x^{1/2} \log^2 x)$;
(ii) $\psi(x) = O(x)$ *if and only if* $\theta(x) = O(x)$, *and in this case* $\psi(x) - \theta(x) = O(x^{1/2} \log x)$.

Proof Since

$$\psi(x) = \sum_{p \le x} \log p + \sum_{p^2 \le x} \log p + \cdots$$

and $k > \log x / \log 2$ implies $x^{1/k} < 2$, we have

$$\psi(x) = \theta(x) + \theta(x^{1/2}) + \cdots + \theta(x^{1/m}),$$

where $m = \lfloor \log x / \log 2 \rfloor$. (As is now usual, we denote by $\lfloor y \rfloor$ the greatest integer $\le y$.) But it is obvious from the definition of $\theta(x)$ that $\theta(x) = O(x \log x)$. Hence

$$\psi(x) - \theta(x) = O\Big(\sum_{2 \le k \le m} x^{1/k} \log x \Big) = O(x^{1/2} \log^2 x).$$

If $\theta(x) = O(x)$ the same argument yields $\psi(x) - \theta(x) = O(x^{1/2} \log x)$ and thus $\psi(x) = O(x)$. It is trivial that $\psi(x) = O(x)$ implies $\theta(x) = O(x)$. □

The proof of Lemma 3 shows also that

$$\psi(x) = \theta(x) + \theta(x^{1/2}) + O(x^{1/3} \log^2 x).$$

Lemma 4 *$\psi(x) = O(x)$ if and only if $\pi(x) = O(x / \log x)$, and then*

$$\pi(x) \log x / x = \psi(x)/x + O(1/\log x).$$

Proof Although their use can easily be avoided, it is more suggestive to use Stieltjes integrals. Suppose first that $\psi(x) = O(x)$. For any $x > 2$ we have

$$\pi(x) = \int_{2-}^{x+} 1/\log t \, d\theta(t)$$

and hence, on integrating by parts,

$$\pi(x) = \theta(x)/\log x + \int_2^x \theta(t)/t \log^2 t \, dt.$$

But

$$\int_2^x \theta(t)/t \log^2 t \, dt = O(x/\log^2 x),$$

since $\theta(t) = O(t)$ and, as we saw in §1,

$$\int_2^x dt/\log^2 t = O(x/\log^2 x).$$

Since

$$\theta(x)/\log x = \psi(x)/\log x + O(x^{1/2}),$$

by Lemma 3, it follows that

$$\pi(x) = \psi(x)/\log x + O(x/\log^2 x).$$

Suppose next that $\pi(x) = O(x/\log x)$. For any $x > 2$ we have

$$\begin{aligned}\theta(x) &= \int_{2-}^{x+} \log t \, d\pi(t) \\ &= \pi(x)\log x - \int_2^x \pi(t)/t \, dt = O(x),\end{aligned}$$

and hence also $\psi(x) = O(x)$, by Lemma 3. □

It follows at once from Lemma 4 that *the prime number theorem*, $\pi(x) \sim x/\log x$, *is equivalent to* $\psi(x) \sim x$.

The method of argument used in Lemma 4 can be carried further. Put

$$\theta(x) = x + R(x), \quad \pi(x) = \int_2^x dt/\log t + Q(x).$$

Subtracting

$$\int_2^x dt/\log t = x/\log x - 2/\log 2 + \int_2^x dt/\log^2 t$$

from

$$\pi(x) = \theta(x)/\log x + \int_2^x \theta(t)/t\log^2 t \, dt,$$

we obtain

$$Q(x) = R(x)/\log x + \int_2^x R(t)/t\log^2 t \, dt + 2/\log 2. \tag{3}_1$$

Also, adding

$$\begin{aligned}\int_2^x \left(\int_2^t du/\log u\right) dt/t &= \int_2^x \left(\int_u^x dt/t\right) du/\log u \\ &= \int_2^x (\log x - \log u) du/\log u \\ &= \log x \int_2^x dt/\log t - x + 2\end{aligned}$$

to

$$\theta(x) = \pi(x)\log x - \int_2^x \pi(t)/t \, dt$$

we obtain

$$R(x) = Q(x)\log x - \int_2^x Q(t)/t \, dt - 2. \tag{3}_2$$

It follows from $(3)_1$–$(3)_2$ that $R(x) = O(x/\log^\alpha x)$ for some $\alpha > 0$ if and only if $Q(x) = O(x/\log^{\alpha+1} x)$. Consequently, by Lemma 3,

$$\psi(x) = x + O(x/\log^\alpha x) \quad \textit{for every } \alpha > 0$$

if and only if

$$\pi(x) = \int_2^x dt/\log t + O(x/\log^\alpha x) \quad \textit{for every } \alpha > 0,$$

and $\pi(x)$ then has the asymptotic expansion

$$\pi(x) \sim \{1 + 1!/\log x + 2!/\log^2 x + \cdots\} x/\log x,$$

the error in breaking off the series after any finite number of terms having the order of magnitude of the first term omitted.

It follows from $(3)_1$–$(3)_2$ also that, *for a given* α *such that* $1/2 \le \alpha < 1$,

$$\psi(x) = x + O(x^\alpha \log^2 x),$$

if and only if

$$\pi(x) = \int_2^x dt/\log t + O(x^\alpha \log x).$$

The definition of $\psi(x)$ can be put in the form

$$\psi(x) = \sum_{n \le x} \Lambda(n),$$

where the *von Mangoldt function* $\Lambda(n)$ is defined by

$$\begin{aligned} \Lambda(n) &= \log p \text{ if } n = p^\alpha \text{ for some prime } p \text{ and some } \alpha > 0, \\ &= 0 \text{ otherwise.} \end{aligned}$$

For any positive integer n we have

$$\log n = \sum_{d|n} \Lambda(d), \tag{4}$$

since if $n = p_1^{\alpha_1} \cdots p_s^{\alpha_s}$ is the factorization of n into powers of distinct primes, then

$$\log n = \sum_{j=1}^{s} \alpha_j \log p_j.$$

3 Proof of the Prime Number Theorem

The *Riemann zeta-function* is defined by

$$\zeta(s) = \sum_{n=1}^{\infty} 1/n^s. \tag{5}$$

This infinite series had already been considered by Euler, Dirichlet and Chebyshev, but Riemann was the first to study it for complex values of s. As customary, we write $s = \sigma + it$, where σ and t are real, and n^{-s} is defined for complex values of s by

$$n^{-s} = e^{-s \log n} = n^{-\sigma}(\cos(t \log n) - i \sin(t \log n)).$$

To show that the series (5) converges in the half-plane $\sigma > 1$ we compare as in §1 the sum with an integral. If $\lfloor x \rfloor$ denotes again the greatest integer $\leq x$, then on integrating by parts we obtain

$$\int_1^N x^{-s} dx - \sum_{n=1}^{N} n^{-s} = \int_{1-}^{N+} x^{-s} d\{x - \lfloor x \rfloor\}$$
$$= -1 + s \int_1^N x^{-s-1} \{x - \lfloor x \rfloor\}\, dx.$$

Since

$$\int_1^N x^{-s} dx = (1 - N^{1-s})/(s-1),$$

by letting $N \to \infty$ we see that $\zeta(s)$ is defined for $\sigma > 1$ and

$$\zeta(s) = 1/(s-1) + 1 - s \int_1^\infty x^{-s-1} \{x - \lfloor x \rfloor\}\, dx.$$

But, since $x - \lfloor x \rfloor$ is bounded, the integral on the right is uniformly convergent in any half-plane $\sigma \geq \delta > 0$. It follows that the definition of $\zeta(s)$ can be extended to the half-plane $\sigma > 0$, so that it is holomorphic there except for a simple pole with residue 1 at $s = 1$.

The connection between the zeta-function and prime numbers is provided by *Euler's product formula*, which may be viewed as an analytic version of the fundamental theorem of arithmetic:

Proposition 5 *$\zeta(s) = \prod_p (1 - p^{-s})^{-1}$ for $\sigma > 1$, where the product is taken over all primes p.*

Proof For $\sigma > 0$ we have

$$(1 - p^{-s})^{-1} = 1 + p^{-s} + p^{-2s} + \cdots.$$

Since each positive integer can be uniquely expressed as a product of powers of distinct primes, it follows that

$$\prod_{p \leq x} (1 - p^{-s})^{-1} = \sum_{n \leq N_x} n^{-s},$$

where N_x is the set of all positive integers, including 1, whose prime factors are all $\leq x$. But N_x contains all positive integers $\leq x$. Hence

$$\left| \zeta(s) - \prod_{p \leq x} (1 - p^{-s})^{-1} \right| \leq \sum_{n > x} n^{-\sigma} \quad \text{for } \sigma > 1,$$

and the sum on the right tends to zero as $x \to \infty$. □

It follows at once from Proposition 5 that $\zeta(s) \neq 0$ for $\sigma > 1$, since the infinite product is convergent and each factor is nonzero.

Proposition 6 $-\zeta'(s)/\zeta(s) = \sum_{n=1}^{\infty} \Lambda(n)/n^s$ *for* $\sigma > 1$, *where* $\Lambda(n)$ *denotes von Mangoldt's function.*

Proof The series $\omega(s) = \sum_{n=1}^{\infty} \Lambda(n) n^{-s}$ converges absolutely and uniformly in any half-plane $\sigma \geq 1 + \varepsilon$, where $\varepsilon > 0$, since

$$0 \leq \Lambda(n) \leq \log n < n^{\varepsilon/2} \quad \text{for all large } n.$$

Hence

$$\begin{aligned} \zeta(s)\omega(s) &= \sum_{m=1}^{\infty} m^{-s} \sum_{k=1}^{\infty} \Lambda(k) k^{-s} \\ &= \sum_{n=1}^{\infty} n^{-s} \sum_{d|n} \Lambda(d). \end{aligned}$$

Since $\sum_{d|n} \Lambda(d) = \log n$, by (4), it follows that

$$\zeta(s)\omega(s) = \sum_{n=1}^{\infty} n^{-s} \log n = -\zeta'(s).$$

Since $\zeta(s) \neq 0$ for $\sigma > 1$, the result follows. However, we can also prove directly that $\zeta(s) \neq 0$ for $\sigma > 1$, and thus make the proof of the prime number theorem independent of Proposition 5.

Obviously if $\zeta(s_0) = 0$ for some s_0 with $\Re s_0 > 1$ then $\zeta'(s_0) = 0$, and it follows by induction from Leibniz' formula for derivatives of a product that $\zeta^{(n)}(s_0) = 0$ for all $n \geq 0$. Since $\zeta(s)$ is holomorphic for $\sigma > 1$ and not identically zero, this is a contradiction. □

Proposition 6 may be restated in terms of Chebyshev's ψ-function:

$$-\zeta'(s)/\zeta(s) = \int_1^{\infty} u^{-s} d\psi(u) = \int_0^{\infty} e^{-sx} d\psi(e^x) \quad \text{for } \sigma > 1. \tag{6}$$

We are going to deduce from (6) that the function $\zeta(s)$ has no zeros on the line $\Re s = 1$. Actually we will prove a more general result:

Proposition 7 *Let* $f(s)$ *be holomorphic in the closed half-plane* $\Re s \geq 1$, *except for a simple pole at* $s = 1$. *If, for* $\Re s > 1$, $f(s) \neq 0$ *and*

$$-f'(s)/f(s) = \int_0^{\infty} e^{-sx} d\phi(x),$$

where $\phi(x)$ *is a nondecreasing function for* $x \geq 0$, *then*

$$f(1 + it) \neq 0 \quad \textit{for every real } t \neq 0.$$

Proof Put $s = \sigma + it$, where σ and t are real, and let

$$g(\sigma, t) = -\mathscr{R}\{f'(s)/f(s)\}.$$

Thus

$$g(\sigma, t) = \int_0^\infty e^{-\sigma x} \cos(tx)\, d\phi(x) \quad \text{for } \sigma > 1.$$

Hence, by Schwarz's inequality (Chapter I, §10),

$$\begin{aligned} g(\sigma, t)^2 &\le \int_0^\infty e^{-\sigma x}\, d\phi(x) \int_0^\infty e^{-\sigma x} \cos^2(tx)\, d\phi(x) \\ &= g(\sigma, 0) \int_0^\infty e^{-\sigma x}\{1 + \cos(2tx)\}\, d\phi(x)/2 \\ &= g(\sigma, 0)\{g(\sigma, 0) + g(\sigma, 2t)\}/2. \end{aligned}$$

Since $f(s)$ has a simple pole at $s = 1$, by comparing the Laurent series of $f(s)$ and $f'(s)$ at $s = 1$ (see Chapter I, §5) we see that

$$(\sigma - 1)g(\sigma, 0) \to 1 \quad \text{as } \sigma \to 1+.$$

Similarly if $f(s)$ has a zero of multiplicity $m(t) \ge 0$ at $1 + it$, where $t \ne 0$, then by comparing the Taylor series of $f(s)$ and $f'(s)$ at $s = 1 + it$ we see that

$$(\sigma - 1)g(\sigma, t) \to -m(t) \quad \text{as } \sigma \to 1+.$$

Thus if we multiply the inequality for $g(\sigma, t)^2$ by $(\sigma - 1)^2$ and let $\sigma \to 1+$, we obtain

$$m(t)^2 \le \{1 - m(2t)\}/2 \le 1/2.$$

Therefore, since $m(t)$ is an integer, $m(t) = 0$. □

For $f(s) = \zeta(s)$, Proposition 7 gives the result of Hadamard and de la Vallée Poussin:

Corollary 8 *$\zeta(1 + it) \ne 0$ for every real $t \ne 0$.*

The use of Schwarz's inequality to prove Corollary 8 seems more natural than the usual proof by means of the inequality $3 + 4\cos\theta + \cos 2\theta \ge 0$. It follows from Corollary 8 that $-\zeta'(s)/\zeta(s) - 1/(s - 1)$ is holomorphic in the closed half-plane $\sigma \ge 1$. Hence, by (6), the hypotheses of the following theorem, due to Ikehara (1931), are satisfied with

$$F(s) = -\zeta'(s)/\zeta(s), \quad \phi(x) = \psi(e^x), \quad h = A = 1.$$

Theorem 9 *Let $\phi(x)$ be a nondecreasing function for $x \ge 0$ such that the Laplace transform*

$$F(s) = \int_0^\infty e^{-sx}\, d\phi(x)$$

is defined for $\mathscr{R}s > h$, where $h > 0$. If there exists a constant A and a function $G(s)$, which is continuous in the closed half-plane $\mathscr{R}s \geq h$, such that

$$G(s) = F(s) - Ah/(s-h) \quad \textit{for } \mathscr{R}s > h,$$

then

$$\phi(x) \sim Ae^{hx} \quad \textit{for } x \to +\infty.$$

Proof For each $X > 0$ we have

$$\int_0^X e^{-sx}\,d\phi(x) = e^{-sX}\{\phi(X) - \phi(0)\} + s\int_0^X e^{-sx}\{\phi(x) - \phi(0)\}\,dx.$$

For real $s = \rho > h$ both terms on the right are nonnegative and the integral on the left has a finite limit as $X \to \infty$. Hence $e^{-\rho X}\phi(X)$ is a bounded function of X for each $\rho > h$. It follows that if $\mathscr{R}s > h$ we can let $X \to \infty$ in the last displayed equation, obtaining

$$F(s) = s\int_0^\infty e^{-sx}\{\phi(x) - \phi(0)\}\,dx \quad \text{for } \mathscr{R}s > h.$$

Hence

$$[G(s) - A]/s = F(s)/s - A/(s-h) = \int_0^\infty e^{-(s-h)x}\{\alpha(x) - A\}\,dx,$$

where $\alpha(x) = e^{-hx}\{\phi(x) - \phi(0)\}$. Thus we will prove the theorem if we prove the following statement:

Let $\alpha(x)$ be a nonnegative function for $x \geq 0$ such that

$$g(s) = \int_0^\infty e^{-sx}\{\alpha(x) - A\}\,dx,$$

where $s = \sigma + it$, is defined for every $\sigma > 0$ and the limit

$$\gamma(t) = \lim_{\sigma \to +0} g(s)$$

exists uniformly on any finite interval $-T \leq t \leq T$. If, for some $h > 0$, $e^{hx}\alpha(x)$ is a nondecreasing function, then

$$\lim_{x\to\infty} \alpha(x) = A.$$

In the proof of this statement we will use the fact that the Fourier transform

$$\hat{k}(u) = \int_{-\infty}^{\infty} e^{iut}k(t)dt$$

of the function

$$k(t) = 1 - |t| \text{ for } |t| \le 1, \ = 0 \text{ for } |t| \ge 1,$$

has the properties

$$\hat{k}(u) \ge 0 \quad \text{for} -\infty < u < \infty, \quad C := \int_{-\infty}^{\infty} \hat{k}(u)du < \infty.$$

Indeed

$$\begin{aligned}\hat{k}(u) &= \int_{-1}^{1} e^{iut}(1-|t|)\,dt \\ &= 2\int_0^1 (1-t)\cos ut\,dt \\ &= 2(1-\cos u)/u^2.\end{aligned}$$

Let ε, λ, y be arbitrary positive numbers. If $s = \varepsilon + i\lambda t$, then

$$\begin{aligned}\lambda\int_{-1}^{1} e^{i\lambda ty}k(t)g(s)\,dt &= \lambda\int_{-1}^{1} e^{i\lambda ty}k(t)\int_0^\infty e^{-\varepsilon x}e^{-i\lambda tx}\{\alpha(x)-A\}\,dxdt \\ &= \lambda\int_0^\infty e^{-\varepsilon x}\{\alpha(x)-A\}\int_{-1}^{1} e^{i\lambda t(y-x)}k(t)\,dtdx \\ &= \lambda\int_0^\infty e^{-\varepsilon x}\alpha(x)\hat{k}(\lambda(y-x))\,dx \\ &\quad - \lambda A\int_0^\infty e^{-\varepsilon x}\hat{k}(\lambda(y-x))\,dx.\end{aligned}$$

When $\varepsilon \to +0$ the left side has the limit

$$\chi(y) := \lambda\int_{-1}^{1} e^{i\lambda ty}k(t)\gamma(\lambda t)\,dt$$

and the second term on the right has the limit

$$\lambda A\int_0^\infty \hat{k}(\lambda(y-x))\,dx.$$

Consequently the first term on the right also has a finite limit. It follows that

$$\lambda\int_0^\infty \alpha(x)\hat{k}(\lambda(y-x))\,dx$$

is finite and is the limit of the first term on the right. Thus

$$\begin{aligned}\chi(y) &= \lambda\int_0^\infty \{\alpha(x)-A\}\hat{k}(\lambda(y-x))\,dx \\ &= \int_{-\infty}^{\lambda y}\{\alpha(y-\upsilon/\lambda)-A\}\hat{k}(\upsilon)\,d\upsilon.\end{aligned}$$

By the 'Riemann–Lebesgue lemma', $\chi(y) \to 0$ as $y \to \infty$. In fact this may be proved in the following way. We have

$$\chi(y) = \int_{-\infty}^{\infty} e^{i\lambda t y} \omega(t)\, dt$$

where

$$\omega(t) = \lambda k(t) \gamma(\lambda t).$$

Changing the variable of integration to $t + \pi/\lambda y$, we obtain

$$\chi(y) = -\int_{-\infty}^{\infty} e^{i\lambda t y} \omega(t + \pi/\lambda y)\, dt.$$

Hence

$$2\chi(y) = \int_{-\infty}^{\infty} e^{i\lambda t y} \{\omega(t) - \omega(t + \pi/\lambda y)\}\, dt$$

and

$$2|\chi(y)| \le \int_{-\infty}^{\infty} |\omega(t) - \omega(t + \pi/\lambda y)|\, dt.$$

Since $\omega(t)$ is continuous and vanishes outside a finite interval, it follows that $\chi(y) \to 0$ as $y \to \infty$.

Since

$$\int_{-\infty}^{\lambda y} \hat{k}(v)\, dv \to C \quad \text{as } y \to \infty,$$

we deduce that

$$\lim_{y\to\infty} = \int_{-\infty}^{\lambda y} \alpha(y - v/\lambda)\, \hat{k}(v)\, dv = AC \quad \text{for every } \lambda > 0.$$

We now make use of the fact that $e^{hx}\alpha(x)$ is a nondecreasing function. Choose any $\delta \in (0, 1)$. If $y = x + \delta$, where $x \ge 0$, then for $|v| \le \lambda\delta$

$$\alpha(y - v/\lambda) \ge e^{-h(\delta - v/\lambda)} \alpha(x) \ge e^{-2h\delta} \alpha(x)$$

and hence

$$\int_{-\infty}^{\lambda y} \alpha(y - v/\lambda) \hat{k}(v)\, dv \ge e^{-2h\delta} \alpha(x) \int_{-\lambda\delta}^{\lambda\delta} \hat{k}(v)\, dv.$$

We can choose $\lambda = \lambda(\delta)$ so large that the integral on the right exceeds $(1 - \delta)C$. Then, letting $x \to \infty$ we obtain

$$AC \ge e^{-2h\delta}(1 - \delta) C \overline{\lim_{x\to\infty}} \alpha(x).$$

Since this holds for arbitrarily small $\delta > 0$, it follows that

$$\overline{\lim_{x\to\infty}}\ \alpha(x) \le A.$$

Thus there exists a positive constant M such that

$$0 \le \alpha(x) \le M \quad \text{for all } x \ge 0.$$

On the other hand, if $y = x - \delta$, where $x \ge \delta$, then for $|v| \le \lambda\delta$

$$\alpha(y - v/\lambda) \le e^{h(\delta+v/\lambda)}\alpha(x) \le e^{2h\delta}\alpha(x)$$

and hence

$$\int_{-\infty}^{\lambda y} \alpha(y - v/\lambda)\hat{k}(v)\,dv \le e^{2h\delta}\alpha(x)\int_{-\lambda\delta}^{\lambda\delta}\hat{k}(v)\,dv + M\int_{|v|\ge\lambda\delta}\hat{k}(v)\,dv.$$

We can choose $\lambda = \lambda(\delta)$ so large that the second term on the right is less than δC. Then, letting $x \to \infty$ we obtain

$$AC \le e^{2h\delta}C\ \underline{\lim_{x\to\infty}}\ \alpha(x) + \delta C.$$

Since this holds for arbitrarily small $\delta > 0$, it follows that

$$A \le \underline{\lim_{x\to\infty}}\ \alpha(x).$$

Combining this with the inequality of the previous paragraph, we conclude that $\lim_{x\to\infty}\alpha(x) = A$. □

Applying Theorem 9 to the special case mentioned before the statement of the theorem, we obtain $\psi(e^x) \sim e^x$. As we have already seen in §2, this is equivalent to the prime number theorem.

4 The Riemann Hypothesis

In his celebrated paper on the distribution of prime numbers Riemann (1859) proved only two results. He showed that the definition of $\zeta(s)$ can be extended to the whole complex plane, so that $\zeta(s) - 1/(s-1)$ is everywhere holomorphic, and he proved that the values of $\zeta(s)$ and $\zeta(1-s)$ are connected by a certain functional equation. This functional equation will now be derived by one of the two methods which Riemann himself used. It is based on a remarkable identity which Jacobi (1829) used in his treatise on elliptic functions.

Proposition 10 *For any $t, y \in \mathbb{R}$ with $y > 0$,*

$$\sum_{n=-\infty}^{\infty} e^{-(t+n)^2\pi y} = y^{-1/2}\sum_{n=-\infty}^{\infty} e^{-n^2\pi/y}e^{2\pi i n t}. \tag{7}$$

In particular,

$$\sum_{n=-\infty}^{\infty} e^{-n^2\pi y} = y^{-1/2}\sum_{n=-\infty}^{\infty} e^{-n^2\pi/y}. \tag{8}$$

Proof Put $f(v) = e^{-v^2\pi y}$ and let

$$g(u) = \int_{-\infty}^{\infty} f(v)e^{-2\pi iuv}\,dv$$

be the Fourier transform of $f(v)$. We are going to show that

$$\sum_{n=-\infty}^{\infty} f(v+n) = \sum_{n=-\infty}^{\infty} g(n)e^{2\pi inv}.$$

Let

$$F(v) = \sum_{n=-\infty}^{\infty} f(v+n).$$

This infinite series is uniformly convergent for $0 \le v \le 1$, and so also is the series obtained by term by term differentiation. Hence $F(v)$ is a continuously differentiable function. Consequently, since it is periodic with period 1, it is the sum of its own Fourier series:

$$F(v) = \sum_{m=-\infty}^{\infty} c_m e^{2\pi imv},$$

where

$$c_m = \int_0^1 F(v)e^{-2\pi imv}\,dv.$$

We can evaluate c_m by term by term integration:

$$\begin{aligned} c_m = \sum_{n=-\infty}^{\infty} \int_0^1 f(v+n)e^{-2\pi imv}\,dv &= \sum_{n=-\infty}^{\infty} \int_n^{n+1} f(v)e^{-2\pi imv}\,dv \\ &= \int_{-\infty}^{\infty} f(v)e^{-2\pi imv}\,dv = g(m). \end{aligned}$$

The argument up to this point is an instance of *Poisson's summation formula*. To evaluate $g(u)$ in the case $f(v) = e^{-v^2\pi y}$ we differentiate with respect to u and integrate by parts, obtaining

$$\begin{aligned} g'(u) &= -2\pi i \int_{-\infty}^{\infty} e^{-v^2\pi y} v e^{-2\pi iuv}\,dv \\ &= (i/y) \int_{-\infty}^{\infty} e^{-2\pi iuv}\,de^{-v^2\pi y} \\ &= -(i/y) \int_{-\infty}^{\infty} e^{-v^2\pi y}\,de^{-2\pi iuv} \\ &= -(2\pi u/y)g(u). \end{aligned}$$

The solution of this first order linear differential equation is

$$g(u) = g(0)e^{-\pi u^2/y}.$$

Moreover

$$g(0) = \int_{-\infty}^{\infty} e^{-v^2\pi y} dv = (\pi y)^{-1/2} J,$$

where

$$J = \int_{-\infty}^{\infty} e^{-v^2} dv.$$

Thus we have proved that

$$\sum_{n=-\infty}^{\infty} e^{-(v+n)^2\pi y} = (\pi y)^{-1/2} J \sum_{n=-\infty}^{\infty} e^{-n^2\pi/y} e^{2\pi i n v}.$$

Substituting $v = 0$, $y = 1$, we obtain $J = \pi^{1/2}$. □

The *theta function*

$$\vartheta(x) = \sum_{n=-\infty}^{\infty} e^{-n^2\pi x} \quad (x > 0)$$

arises not only in the theory of elliptic functions, as we will see in Chapter XII, but also in problems of heat conduction and statistical mechanics. The transformation law

$$\vartheta(x) = x^{-1/2}\vartheta(1/x)$$

is very useful for computational purposes since, when x is small, the series for $\vartheta(x)$ converges extremely slowly but the series for $\vartheta(1/x)$ converges extremely rapidly.

Since the functional equation of Riemann's zeta function involves Euler's *gamma function*, we summarize here the main properties of the latter. Euler (1729) defined his function $\Gamma(z)$ by

$$1/\Gamma(z) = \lim_{n\to\infty} z(z+1)\cdots(z+n)/n!\, n^z,$$

where $n^z = \exp(z \log n)$ and the limit exists for every $z \in \mathbb{C}$. It follows from the definition that $1/\Gamma(z)$ is everywhere holomorphic and that its only zeros are simple zeros at the points $z = 0, -1, -2, \ldots$. Moreover $\Gamma(1) = 1$ and

$$\Gamma(z+1) = z\Gamma(z).$$

Hence $\Gamma(n+1) = n!$ for any positive integer n. By putting $\Gamma(z+1) = z!$ the definition of the factorial function may be extended to any $z \in \mathbb{C}$ which is not a negative integer. Wielandt (1939) has characterized $\Gamma(z)$ as the only solution of the functional equation

$$F(z+1) = zF(z)$$

with $F(1) = 1$ which is holomorphic in the half-plane $\mathscr{R}z > 0$ and bounded for $1 < \mathscr{R}z < 2$.

It follows from the definition of $\Gamma(z)$ and the product formula for the sine function that

$$\Gamma(z)\Gamma(1-z) = \pi / \sin \pi z.$$

Many definite integrals may be evaluated in terms of the gamma function. By repeated integration by parts it may be seen that, if $\mathscr{R}z > 0$ and $n \in \mathbb{N}$, then

$$n! n^z / z(z+1)\cdots(z+n) = \int_0^n (1 - t/n)^n t^{z-1} dt,$$

where $t^{z-1} = \exp\{(z-1)\log t\}$. Letting $n \to \infty$, we obtain the integral representation

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt \quad \text{for } \mathscr{R}z > 0. \tag{9}$$

It follows that $\Gamma(1/2) = \pi^{1/2}$, since

$$\int_0^\infty e^{-t} t^{-1/2} dt = \int_{-\infty}^\infty e^{-v^2} dv = \pi^{1/2},$$

by the proof of Proposition 10. It was already shown by Euler (1730) that

$$\mathrm{B}(x, y) := \int_0^1 t^{x-1}(1-t)^{y-1} dt = \Gamma(x)\Gamma(y)/\Gamma(x+y),$$

the relation holding for $\mathscr{R}x > 0$ and $\mathscr{R}y > 0$. The unit ball in $\mathbb{R}^n$ has volume $\kappa_n := \pi^{n/2}/(n/2)!$ and surface content $n\kappa_n$. *Stirling's formula*, $n! \approx (n/e)^n \sqrt{2\pi n}$, follows at once from the integral representation

$$\log \Gamma(z) = (z - 1/2)\log z - z + (1/2)\log 2\pi - \int_0^\infty (t - \lfloor t \rfloor - 1/2)(z+t)^{-1} dt,$$

valid for any $z \in \mathbb{C}$ which is not zero or a negative integer. Euler's constant

$$\gamma = \lim_{n\to\infty} (1 + 1/2 + 1/3 + \cdots + 1/n - \log n) \approx 0.5772157$$

may also be defined by $\gamma = -\Gamma'(1)$.

We now return to the Riemann zeta function.

Proposition 11 *The function $Z(s) = \pi^{-s/2}\Gamma(s/2)\zeta(s)$ satisfies the functional equation*

$$Z(s) = Z(1-s) \text{ for } 0 < \sigma < 1.$$

Proof From the representation (9) of the gamma function we obtain, for $\sigma > 0$ and $n \geq 1$,

$$\int_0^\infty x^{s/2-1}e^{-n^2\pi x}dx = \pi^{-s/2}\Gamma(s/2)n^{-s}.$$

Hence, if $\sigma > 1$,

$$\begin{aligned} Z(s) &= \sum_{n=1}^{\infty}\int_0^\infty x^{s/2-1}e^{-n^2\pi x}dx \\ &= \int_0^\infty x^{s/2-1}\phi(x)dx, \end{aligned}$$

where

$$\phi(x) = \sum_{n=1}^{\infty} e^{-n^2\pi x}.$$

By Proposition 10,

$$2\phi(x) + 1 = x^{-1/2}[2\phi(1/x) + 1].$$

Hence

$$\begin{aligned} Z(s) &= \int_1^\infty x^{s/2-1}\phi(x)\,dx + \int_0^1 x^{s/2-1}\{x^{-1/2}\phi(1/x) + (1/2)x^{-1/2} - 1/2\}\,dx \\ &= \int_1^\infty x^{s/2-1}\phi(x)\,dx + \int_0^1 x^{s/2-3/2}\phi(1/x)\,dx + 1/(s-1) - 1/s \\ &= \int_1^\infty (x^{s/2-1} + x^{-s/2-1/2})\phi(x)\,dx + 1/s(s-1). \end{aligned}$$

The integral on the right is convergent for all s and thus provides the analytic continuation of $Z(s)$ to the whole plane. Moreover the right side is unchanged if s is replaced by $1 - s$. □

The function $Z(s)$ in Proposition 11 is occasionally called the *completed* zeta function. In its product representation

$$Z(s) = \pi^{-s/2}\Gamma(s/2)\Pi_p(1 - p^{-s})^{-1}$$

it makes sense to regard $\pi^{-s/2}\Gamma(s/2)$ as an Euler factor at ∞, complementing the Euler factors $(1 - p^{-s})^{-1}$ at the primes p.

It follows from Proposition 11 and the previously stated properties of the gamma function that the definition of $\zeta(s)$ may be extended to the whole complex plane, so that $\zeta(s) - 1/(s-1)$ is everywhere holomorphic and $\zeta(s) = 0$ if $s = -2, -4, -6, \ldots$. Since $\zeta(s) \neq 0$ for $\sigma \geq 1$ and $\zeta(0) = -1/2$, the functional equation shows that these 'trivial' zeros of $\zeta(s)$ are its only zeros in the half-plane $\sigma \leq 0$. Hence all 'nontrivial' zeros of $\zeta(s)$ lie in the strip $0 < \sigma < 1$ and are symmetrically situated with respect to the line $\sigma = 1/2$. The famous *Riemann hypothesis* asserts that all zeros in this strip actually lie on the line $\sigma = 1/2$.

Since $\zeta(\bar{s}) = \overline{\zeta(s)}$, the zeros of $\zeta(s)$ are also symmetric with respect to the real axis. Furthermore $\zeta(s)$ has no real zeros in the strip $0 < \sigma < 1$, since

$$(1 - 2^{-1-\sigma})\zeta(\sigma) = (1 - 2^{-\sigma}) + (3^{-\sigma} - 4^{-\sigma}) + \cdots > 0 \quad \text{for } 0 < \sigma < 1.$$

It has been verified by van de Lune *et al.* (1986), with the aid of a supercomputer, that the 1.5×10^9 zeros of $\zeta(s)$ in the rectangle $0 < \sigma < 1, 0 < t < T$, where $T = 545439823.215$, are all simple and lie on the line $\sigma = 1/2$.

The location of the zeros of $\zeta(s)$ is intimately connected with the asymptotic behaviour of $\pi(x)$. Let α^* denote the least upper bound of the real parts of all zeros of $\zeta(s)$. Then $1/2 \leq \alpha^* \leq 1$, since it is known that $\zeta(s)$ does have zeros in the strip $0 < \sigma < 1$, and the Riemann hypothesis is equivalent to $\alpha^* = 1/2$. It was shown by von Koch (1901) that

$$\psi(x) = x + O(x^{\alpha^*} \log^2 x)$$

and hence

$$\pi(x) = Li(x) + O(x^{\alpha^*} \log x).$$

(Actually von Koch assumed $\alpha^* = 1/2$, but his argument can be extended without difficulty.) It should be noted that these estimates are of interest only if $\alpha^* < 1$.

On the other hand if, for some α such that $0 < \alpha < 1$,

$$\pi(x) = Li(x) + O(x^{\alpha} \log x),$$

then

$$\theta(x) = x + O(x^{\alpha} \log^2 x).$$

By the remark after the proof of Lemma 3, it follows that

$$\psi(x) = x + x^{1/2} + O(x^{\alpha} \log^2 x) + O(x^{1/3} \log^2 x).$$

But for $\sigma > 1$ we have

$$-\zeta'(s)/\zeta(s) = \int_1^{\infty} x^{-s} d\psi(x) = s \int_1^{\infty} \psi(x) x^{-s-1} dx$$

and hence

$$-\zeta'(s)/\zeta(s) - s/(s-1) - s/(s-1/2) = s \int_1^{\infty} \{\psi(x) - x - x^{1/2}\} x^{-s-1} dx.$$

The integral on the right is uniformly convergent in the half-plane $\sigma \geq \varepsilon + \max(\alpha, 1/3)$, for any $\varepsilon > 0$, and represents there a holomorphic function. It follows that $1/2 \leq \alpha^* \leq \max(\alpha, 1/3)$. Consequently $\alpha^* \leq \alpha$ and $\psi(x) = x + O(x^{\alpha} \log^2 x)$.

Combining this with von Koch's result, we see that the Riemann hypothesis is equivalent to

$$\pi(x) = Li(x) + O(x^{1/2} \log x)$$

and to

$$\psi(x) = x + O(x^{1/2}\log^2 x).$$

Since it is still not known if $\alpha^* < 1$, the error terms here are substantially smaller than any that have actually been established.

It has been shown by Cramér (1922) that

$$(\log x)^{-1}\int_2^x (\psi(t)/t - 1)^2 dt$$

has a finite limit as $x \to \infty$ if the Riemann hypothesis holds, and is unbounded if it does not. Similarly, for each $\alpha < 1$,

$$x^{-2(1-\alpha)}\int_2^x (\psi(t) - t)^2 t^{-2\alpha} dt$$

is bounded but does not have a finite limit as $x \to \infty$ if the Riemann hypothesis holds, and is unbounded otherwise.

For all values of x listed in Table 1 we have $\pi(x) < Li(x)$, and at one time it was conjectured that this inequality holds for all $x > 0$. However, Littlewood (1914) disproved the conjecture by showing that there exists a constant $c > 0$ such that

$$\pi(x_n) - Li(x_n) > cx_n^{1/2}\log\log\log x_n/\log x_n$$

for some sequence $x_n \to \infty$ and

$$\pi(\xi_n) - Li(\xi_n) < -c\xi_n^{1/2}\log\log\log \xi_n/\log \xi_n$$

for some sequence $\xi_n \to \infty$. This is a quite remarkable result, since no actual value of x is known for which $\pi(x) > Li(x)$. However, it is known that $\pi(x) > Li(x)$ for some x between 1.398201×10^{316} and 1.398244×10^{316}.

In this connection it may be noted that Rosser and Schoenfeld (1962) have shown that $\pi(x) > x/\log x$ for all $x \geq 17$. It had previously been shown by Rosser (1939) that $p_n > n\log n$ for all $n \geq 1$.

Not content with not being able to prove the Riemann hypothesis, Montgomery (1973) has assumed it and made a further conjecture. For given $\beta > 0$, let $N_T(\beta)$ be the number of zeros $1/2 + i\gamma$, $1/2 + i\gamma'$ of $\zeta(s)$ with $0 < \gamma' < \gamma \leq T$ such that

$$\gamma - \gamma' \leq 2\pi\beta/\log T.$$

Montgomery's conjecture is that, for each fixed $\beta > 0$,

$$N_T(\beta) \sim (T/2\pi)\log T\int_0^\beta \{1 - (\sin\pi u/\pi u)^2\}\,du \quad \text{as } T \to \infty.$$

Goldston (1988) has shown that this is equivalent to

$$\int_1^{T^\beta} \{\psi(x + x/T) - \psi(x) - x/T\}^2 x^{-2} dx \sim (\beta - 1/2)\log^2 T/T \quad \text{as } T \to \infty,$$

for each fixed $\beta \geq 1$, where $\psi(x)$ is Chebyshev's function.

In the language of physics Montgomery's conjecture says that $1 - (\sin \pi u/\pi u)^2$ is the *pair correlation function* of the zeros of $\zeta(s)$. Dyson pointed out that this is also the pair correlation function of the normalized eigenvalues of a random $N \times N$ Hermitian matrix in the limit $N \to \infty$. A great deal more is known about this so-called *Gaussian unitary ensemble*, which Wigner (1955) used to model the statistical properties of the spectra of complex nuclei. For example, if the eigenvalues are normalized so that the average difference between consecutive eigenvalues is 1, then the probability that the difference between an eigenvalue and the least eigenvalue greater than it does not exceed β converges as $N \to \infty$ to

$$\int_0^\beta p(u)\, du,$$

where the density function $p(u)$ can be explicitly specified.

It has been further conjectured that the spacings of the normalized zeros of the zeta-function have the same distribution. To make this precise, let the zeros $1/2 + i\gamma_n$ of $\zeta(s)$ with $\gamma_n > 0$ be numbered so that

$$\gamma_1 \le \gamma_2 \le \cdots .$$

Since it is known that the number of γ's in an interval $[T, T+1]$ is asymptotic to $(\log T)/2\pi$ as $T \to \infty$, we put

$$\tilde{\gamma}_n = (\gamma_n \log \gamma_n)/2\pi,$$

so that the average difference between consecutive $\tilde{\gamma}_n$ is 1. If $\delta_n = \tilde{\gamma}_{n+1} - \tilde{\gamma}_n$, and if $v_N(\beta)$ is the number of $\delta_n \le \beta$ with $n \le N$, then the conjecture is that for each $\beta > 0$

$$v_N(\beta)/N \to \int_0^\beta p(u)\, du \quad \text{as } N \to \infty.$$

This nearest neighbour conjecture and the Montgomery pair correlation conjecture have been extensively tested by Odlyzko (1987/9) with the aid of a supercomputer. There is good agreement between the conjectures and the numerical results.

5 Generalizations and Analogues

The prime number theorem may be generalized to any algebraic number field in the following way. Let K be an algebraic number field, i.e. a finite extension of the field $\mathbb{Q}$ of rational numbers. Let R be the ring of all algebraic integers in K, $\mathscr{I}$ the set of all nonzero ideals of R, and $\mathscr{P}$ the subset of prime ideals. For any $A \in \mathscr{I}$, the quotient ring R/A is finite; its cardinality will be denoted by $|A|$ and called the *norm* of A.

It may be shown that the *Dedekind zeta-function*

$$\zeta_K(s) = \sum_{A \in \mathscr{I}} |A|^{-s}$$

is defined for $\mathscr{R}s > 1$ and that the product formula

$$\zeta_K(s) = \prod_{P \in \mathscr{P}} (1 - |P|^{-s})^{-1}$$

holds in this open half-plane. Furthermore the definition of $\zeta_K(s)$ may be extended so that it is nonzero and holomorphic in the closed half-plane $\mathscr{R}s \geq 1$, except for a simple pole at $s = 1$. By applying Ikehara's theorem we can then obtain the *prime ideal theorem*, which was first proved by Landau (1903):

$$\pi_K(x) \sim x/\log x,$$

where $\pi_K(x)$ denotes the number of prime ideals of R with norm $\leq x$.

It was shown by Hecke (1917) that the definition of the Dedekind zeta-function $\zeta_K(s)$ may also be extended so that it is holomorphic in the whole complex plane, except for the simple pole at $s = 1$, and so that, for some constant $A > 0$ and non-negative integers r_1, r_2 (which can all be explicitly described in terms of the structure of the algebraic number field K),

$$Z_K(s) = A\Gamma(s/2)^{r_1}\Gamma(s)^{r_2}\zeta_K(s)$$

satisfies the functional equation

$$Z_K(s) = Z_K(1-s).$$

The *extended Riemann hypothesis* asserts that, for every algebraic number field K,

$$\zeta_K(s) \neq 0 \quad \text{for } \mathscr{R}s > 1/2.$$

The numerical evidence for the extended Riemann hypothesis is favourable, although in the nature of things it cannot be tested as extensively as the ordinary Riemann hypothesis. The extended Riemann hypothesis implies error bounds for the prime ideal theorem of the same order as those which the ordinary Riemann hypothesis implies for the prime number theorem. However, it also has many other consequences. We mention only two.

It has been shown by Bach (1990), making precise an earlier result of Ankeny (1952), that if the extended Riemann hypothesis holds then, for each prime p, there is a quadratic non-residue a mod p with $a < 2\log^2 p$. Thus we do not have to search far in order to find a quadratic non-residue, or to disprove the extended Riemann hypothesis.

It will be recalled from Chapter II that if p is a prime and a an integer not divisible by p, then $a^{p-1} \equiv 1 \bmod p$. For each prime p there exists a *primitive root*, i.e. an integer a such that $a^k \not\equiv 1 \bmod p$ for $1 \leq k < p-1$. It is easily seen that an even square is never a primitive root, that an odd square (including 1) is a primitive root only for the prime $p = 2$, and that -1 is a primitive root only for the primes $p = 2, 3$.

Assuming the extended Riemann hypothesis, Hooley (1967) has proved a famous conjecture of Artin (1927): if the integer a is not a square or -1, then there exist infinitely many primes p for which a is a primitive root. Moreover, if $N_a(x)$ denotes the number of primes $p \leq x$ for which a is a primitive root, then

$$N_a(x) \sim A_a x/\log x \quad \text{for } x \to \infty,$$

where A_a is a positive constant which can be explicitly described. (The expression for A_a which Artin conjectured requires modification in some cases.)

There are also analogues for function fields of these results for number fields. Let K be an arbitrary field. A *field of algebraic functions of one variable over* K is a field L which satisfies the following conditions:

(i) $K \subseteq L$,

(ii) L contains an element v which is *transcendental* over K, i.e. v satisfies no monic polynomial equation

$$u^n + a_1 u^{n-1} + \cdots + a_n = 0$$

with coefficients $a_j \in K$,

(iii) L is a *finite extension* of the field $K(v)$ of rational functions of v with coeffients from K, i.e. L is finite-dimensional as a vector space over $K(v)$.

Let R be a ring with $K \subseteq R \subset L$ such that $x \in L \backslash R$ implies $x^{-1} \in R$. Then the set P of all $a \in R$ such that $a = 0$ or $a^{-1} \notin R$ is an ideal of R, and actually the unique maximal ideal of R. Hence the quotient ring R/P is a field. Since R is the set of all $x \in L$ such that $xP \subseteq P$, it is uniquely determined by P. The ideal P will be called a *prime divisor* of the field L and R/P its *residue field*. It may be shown that the residue field R/P is a finite extension of (a field isomorphic to) K.

An arbitrary *divisor* of the field L is a formal product $A = \prod_P P^{v_P}$ over all prime divisors P of L, where the exponents v_P are integers only finitely many of which are nonzero. The divisor is *integral* if $v_P \geq 0$ for all P.

The set K' of all elements of L which satisfy monic polynomial equations with coefficients from K is a subfield containing K, and L is also a field of algebraic functions of one variable over K'. It is easily shown that no element of $L \backslash R$ satisfies a monic polynomial equation with coefficients from R. Consequently $K' \subseteq R$ and the notion of prime divisor is the same whether we consider L to be over K or over K'. Since $(K')' = K'$, we may assume from the outset that $K' = K$. The elements of K will then be called *constants* and the elements of L *functions*.

Suppose now that the field of constants K is a finite field $\mathbb{F}_q$ containing q elements. We define the *norm* $N(P)$ of a prime divisor P to be the cardinality of the associated residue field R/P and the norm of an integral divisor $A = \prod_P P^{v_P}$ to be

$$N(A) = \prod_P N(P)^{v_P}.$$

It may be shown that, for each positive integer m, there exist only finitely many prime divisors of norm q^m. Moreover, for $\mathscr{R}s > 1$ the zeta-function of L can be defined by

$$\zeta_L(s) = \sum_A N(A)^{-s},$$

where the sum is over all integral divisors of L, and then

$$\zeta_L(s) = \prod_P (1 - N(P)^{-s})^{-1},$$

where the product is over all prime divisors of L.

This seems quite similar to the number field case, but the function field case is actually simpler. F.K. Schmidt (1931) deduced from the Riemann–Roch theorem that there exists a polynomial $p(u)$ of even degree $2g$, with integer coefficients and constant term 1, such that

$$\zeta_L(s) = p(q^{-s})/(1-q^{-s})(1-q^{1-s}),$$

and that the zeta-function satisfies the functional equation

$$q^{(g-1)s}\zeta_L(s) = q^{(g-1)(1-s)}\zeta_L(1-s).$$

The non-negative integer g is the *genus* of the field of algebraic functions.

The analogue of the Riemann hypothesis, that all zeros of $\zeta_L(s)$ lie on the line $\mathscr{R}s = 1/2$, is equivalent to the statement that all zeros of the polynomial $p(u)$ have absolute value $q^{-1/2}$, or that the number N of prime divisors with norm q satisfies the inequality

$$|N - (q+1)| \le 2gq^{1/2}.$$

This analogue has been *proved* by Weil (1948). A simpler proof has been given by Bombieri (1974), using ideas of Stepanov (1969).

The theory of function fields can also be given a geometric formulation. The prime divisors of a function field L with field of constants K can be regarded as the points of a non-singular projective curve over K, and vice versa. Weil (1949) conjectured far-reaching generalizations of the preceding results for curves over a finite field to algebraic varieties of higher dimension.

Let V be a nonsingular projective variety of dimension d, defined by homogeneous polynomials with coefficients in $\mathbb{Z}$. For any prime p, let V_p be the (possibly singular) variety defined by reducing the coefficients mod p and consider the formal power series

$$Z_p(T) := \exp\left(\sum_{n\ge 1} N_n(p)T^n/n\right),$$

where $N_n(p)$ denotes the number of points of V_p defined over the finite field $\mathbb{F}_{p^n}$. Weil conjectured that, if V_p is a nonsingular projective variety of dimension d over $\mathbb{F}_p$, then

(i) $Z_p(T)$ is a rational function of T,
(ii) $Z_p(1/p^dT) = \pm p^{de/2}T^eZ_p(T)$ for some integer e,
(iii) $Z_p(T)$ has a factorization of the form

$$Z_p(T) = P_1(T)\cdots P_{2d-1}(T)/P_0(T)P_2(T)\cdots P_{2d}(T),$$

where $P_0(T) = 1-T$, $P_{2d}(T) = 1-p^dT$ and $P_j(T) \in \mathbb{Z}[T]$ $(0 < j < 2d)$,
(iv) $P_j(T) = \prod_{k=1}^{b_j}(1-\alpha_{jk}T)$, where $|\alpha_{jk}| = p^{j/2}$ for $1 \le k \le b_j$, $(0 < j < 2d)$.

The Weil conjectures have a topological significance, since the integer e in (ii) is the Euler characteristic of the original variety V, regarded as a complex manifold, and b_j in (iv) is its j-th Betti number.

Conjecture (i) was proved by Dwork (1960). The remaining conjectures were proved by Deligne (1974), using ideas of Grothendieck. The most difficult part is the proof that $|\alpha_{jk}| = p^{j/2}$ (the Riemann hypothesis for varieties over finite fields). Deligne's proof is a major achievement of 20th century mathematics, but unfortunately of a different order of difficulty than anything which will be proved here.

An analogue for function fields of Artin's primitive root conjecture was already proved by Bilharz (1937), assuming the Riemann hypothesis for this case. Function fields have been used by Goppa (1981) to construct linear codes. Good codes are obtained when the number of prime divisors is large compared to the genus, and this can be guaranteed by means of the Riemann 'hypothesis'.

Carlitz and Uchiyama (1957) used the Riemann hypothesis for function fields to obtain useful estimates for exponential sums in one variable, and Deligne (1977) showed that these estimates could be extended to exponential sums in several variables. Let $\mathbb{F}_p$ be the field of p elements, where p is a prime, and let $f \in \mathbb{F}_p[u_1, \ldots, u_n]$ be a polynomial in n variables of degree $d \geq 1$ with coefficients from $\mathbb{F}_p$ which is not of the form $g^p - g + b$, where $b \in \mathbb{F}_p$ and $g \in \mathbb{F}_p[u_1, \ldots, u_n]$. (This condition is certainly satisfied if $d < p$.) Then

$$\left| \sum_{x_1, \ldots, x_n \in \mathbb{F}_p} e^{2\pi i f(x_1, \ldots, x_n)/p} \right| \leq (d-1)p^{n-1/2}.$$

We mention one more application of the Weil conjectures. *Ramanujan's tau-function* is defined by

$$q \prod_{n=1}^{\infty} (1-q^n)^{24} = \sum_{n=1}^{\infty} \tau(n) q^n.$$

It was conjectured by Ramanujan (1916), and proved by Mordell (1920), that

$$\sum_{n=1}^{\infty} \tau(n)/n^s = \prod_p (1 - \tau(p)p^{-s} + p^{11-2s})^{-1},$$

where the product is over all primes p. Ramanujan additionally conjectured that $|\tau(p)| \leq 2p^{11/2}$ for all p, and Deligne (1968/9) showed that this was a consequence of the (at that time unproven) Weil conjectures.

The prime number theorem also has an interesting analogue in the theory of dynamical systems. Let M be a compact Riemannian manifold with negative sectional curvatures, and let $N(T)$ denote the number of different (oriented) closed geodesics on M of length $\leq T$. It was first shown by Margulis (1970) that

$$N(T) \sim e^{hT}/hT \quad \text{as } T \to \infty,$$

where the positive constant h is the topological entropy of the associated geodesic flow.

Although much of the detail is specific to the problem, a proof may be given which has the same structure as the proof in §3 of the prime number theorem. If P is an

arbitrary closed orbit of the geodesic flow and $\lambda(P)$ its least period, one shows that the zeta-function

$$\zeta_M(s) = \prod_P (1 - e^{-s\lambda(P)})^{-1}$$

is nonzero and holomorphic for $\mathscr{R}s \geq h$, except for a simple pole at $s = h$, and then applies Ikehara's theorem. The study of geodesics on a surface of negative curvature was initiated by Hadamard (1898), but it is unlikely that he realized there was a connection with the prime number theorem which he had proved two years earlier!

6 Alternative Formulations

There is an intimate connection between the *Dirichlet products* considered in §3 of Chapter III and *Dirichlet series*. It is easily seen that if the Dirichlet series

$$f(s) = \sum_{n=1}^{\infty} a(n)/n^s, \quad g(s) = \sum_{n=1}^{\infty} b(n)/n^s,$$

are absolutely convergent for $\mathscr{R}s > \alpha$, then the product $h(s) = f(s)g(s)$ may also be represented by an absolutely convergent Dirichlet series for $\mathscr{R}s > \alpha$:

$$h(s) = \sum_{n=1}^{\infty} c(n)/n^s,$$

where $c = a * b$, i.e.

$$c(n) = \sum_{d|n} a(d)b(n/d) = \sum_{d|n} a(n/d)b(d).$$

This implies, in particular, that for $\mathscr{R}s > 1$

$$\zeta^2(s) = \sum_{n=1}^{\infty} \tau(n)/n^s, \quad \zeta(s-1)\zeta(s) = \sum_{n=1}^{\infty} \sigma(n)/n^s,$$

where as in Chapter III (not as in §5),

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

denote respectively the number of positive divisors of n and the sum of the positive divisors of n. The relation for Euler's phi-function,

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

which was proved in Chapter III, now yields for $\mathscr{R}s > 1$

$$\zeta(s-1)/\zeta(s) = \sum_{n=1}^{\infty} \varphi(n)/n^s.$$

From the property by which we defined the Möbius function we obtain also, for $\mathscr{R}s > 1$,

$$1/\zeta(s) = \sum_{n=1}^{\infty} \mu(n)/n^s.$$

In view of this relation it is not surprising that the distribution of prime numbers is closely connected with the behaviour of the Möbius function. Put

$$M(x) = \sum_{n \le x} \mu(n).$$

Since $|\mu(n)| \le 1$, it is obvious that $|M(x)| \le \lfloor x \rfloor$ for $x > 0$. The next result is not so obvious:

Proposition 12 *$M(x)/x \to 0$ as $x \to \infty$.*

Proof The function $f(s) := \zeta(s) + 1/\zeta(s)$ is holomorphic for $\sigma \ge 1$, except for a simple pole with residue 1 at $s = 1$. Moreover

$$f(s) = \sum_{n=1}^{\infty} \{1 + \mu(n)\}/n^s = \int_{1-}^{\infty} x^{-s} d\phi(x) \text{ for } \sigma > 1,$$

where $\phi(x) = \lfloor x \rfloor + M(x)$ is a nondecreasing function. Since

$$f(s) = \int_{0-}^{\infty} e^{-su} d\phi(e^u),$$

it follows from Ikehara's Theorem 9 that $\phi(x) \sim x$. □

Proposition 12 is equivalent to the prime number theorem in the sense that either of the relations $M(x) = o(x)$, $\psi(x) \sim x$ may be deduced from the other by elementary (but not trivial) arguments.

The Riemann hypothesis also has an equivalent formulation in terms of the function $M(x)$. Suppose

$$M(x) = O(x^{\alpha}) \quad \text{as } x \to \infty,$$

for some α such that $0 < \alpha < 1$. For $\sigma > 1$ we have

$$1/\zeta(s) = \int_{1-}^{\infty} x^{-s} dM(x) = s \int_{1}^{\infty} x^{-s-1} M(x)\, dx.$$

But for $\sigma > \alpha$ the integral on the right is convergent and defines a holomorphic function. Consequently it is the analytic continuation of $1/\zeta(s)$. Thus if α^* again denotes the least upper bound of all zeros of $\zeta(s)$, then $\alpha \ge \alpha^* \ge 1/2$. On the other hand, Littlewood (1912) showed that

$$M(x) = O(x^{\alpha^*+\varepsilon}) \quad \text{for every } \varepsilon > 0.$$

It follows that *the Riemann hypothesis holds if and only if* $M(x) = O(x^{\alpha})$ *for every* $\alpha > 1/2$.

It has already been mentioned that the first 1.5×10^9 zeros of $\zeta(s)$ on the line $\sigma = 1/2$ are all simple. It is likely that the Riemann hypothesis does not tell the whole story and that all zeros of $\zeta(s)$ on the line $\sigma = 1/2$ are simple. Thus it is of interest that this is guaranteed by a sufficiently sharp bound for $M(x)$. We will show that *if*

$$M(x) = O(x^{1/2}\log^{\alpha} x) \quad \text{as } x \to \infty,$$

for some $\alpha < 1$, *then not only do all nontrivial zeros of* $\zeta(s)$ *lie on the line* $\sigma = 1/2$ *but they are all simple*.

Let $\rho = 1/2 + i\gamma$ be a zero of $\zeta(s)$ of multiplicity $m \geq 1$ and take $s = \rho + h$, where $h > 0$. Then $\sigma = 1/2 + h$ and, since

$$1/\zeta(s) = s\int_1^{\infty} x^{-s-1}M(x)\,dx \text{ for } \sigma > 1/2,$$

we have

$$|1/\zeta(s)| \leq |s|\int_1^{\infty} x^{-\sigma-1}|M(x)|\,dx = O(|s|)\int_1^{\infty} x^{-h-1}\log^{\alpha} x\,dx$$
$$= O(|s|)\int_0^{\infty} e^{-hu}u^{\alpha}du = O(|s|)\Gamma(\alpha+1)/h^{\alpha+1}.$$

Thus $h^{\alpha+1}|1/\zeta(s)|$ is bounded for $h \to +0$ and hence $m \leq \alpha+1$. Since m is an integer and $\alpha < 1$, this implies $m = 1$ and $\alpha \geq 0$.

The prime number theorem, in the form $M(x) = o(x)$, says that asymptotically $\mu(n)$ takes the values $+1$ and -1 with equal probability. By assuming that actually the values $\mu(n)$ asymptotically behave like independent random variables Good and Churchhouse (1968) have been led to two striking conjectures, analogous to the central limit theorem and the law of the iterated logarithm in the theory of probability:

Conjecture A *If* $N(n) \to \infty$ *and* $\log N/\log n \to 0$, *then*

$$P_n\left\{\frac{M(m+N)-M(m)}{(6N/\pi^2)^{1/2}} < t\right\} \to (2\pi)^{-1/2}\int_{-\infty}^{t} e^{-u^2/2}du,$$

where

$$P_n\{f(m) < t\} = \#\{m \leq n : f(m) < t\}/n.$$

Conjecture B

$$\overline{\lim_{x\to\infty}}\, M(x)(2x\log\log x)^{-1/2} = \sqrt{6}/\pi$$
$$= -\underline{\lim_{x\to\infty}}\, M(x)(2x\log\log x)^{-1/2}.$$

By what has been said, Conjecture B implies not only the Riemann hypothesis, but also that the zeros of $\zeta(s)$ are all simple. These probabilistic conjectures provide a more interesting reason than symmetry for believing in the validity of the Riemann hypothesis, but no progress has so far been made towards proving them.

7 Some Further Problems

A prime p is said to be a *twin prime* if $p+2$ is also a prime. For example, 41 is a twin prime since both 41 and 43 are primes. It is still not known if there are infinitely many twin primes. However Brun (1919), using the sieve method which he devised for the purpose, showed that, if infinite, the sum of the reciprocals of all twin primes converges. Since the sum of the reciprocals of all primes diverges, this means that few primes are twin primes.

By a formal application of their circle method Hardy and Littlewood (1923) were led to conjecture that

$$\pi_2(x) \sim L_2(x) \quad \text{for } x \to \infty,$$

where $\pi_2(x)$ denotes the number of twin primes $\leq x$,

$$L_2(x) = 2C_2 \int_2^x dt/\log^2 t$$

and

$$C_2 = \prod_{p\geq 3}(1 - 1/(p-1)^2) = 0.66016181\ldots.$$

This implies that $\pi_2(x)/\pi(x) \sim 2C_2/\log x$. Table 2, adapted from Brent (1975), shows that Hardy and Littlewood's formula agrees well with the facts. Brent also calculates

$$\sum_{\text{twin}\, p\leq 10^{10}} (1/p + 1/(p+2)) = 1.78748\ldots$$

and, using the Hardy–Littlewood formula for the tail, obtains the estimate

$$\sum_{\text{all twin } p} (1/p + 1/(p+2)) = 1.90216\ldots.$$

His calculations have been considerably extended by Nicely (1995).

Table 2.

x	$\pi_2(x)$	$L_2(x)$	$\pi_2(x)/L_2(x)$
10^3	35	46	0.76
10^4	205	214	0.96
10^5	1224	1249	0.980
10^6	8169	8248	0.9904
10^7	58980	58754	1.0038
10^8	440312	440368	0.99987
10^9	3424506	3425308	0.99977
10^{10}	27412679	27411417	1.000046

Besides the twin prime formula many other asymptotic formulae were conjectured by Hardy and Littlewood. Most of them are contained in a general conjecture, which will now be described.

Let $f(t)$ be a polynomial in t of positive degree with integer coefficients. If $f(n)$ is prime for infinitely many positive integers n, then f has positive leading coefficient, f is irreducible over the field $\mathbb{Q}$ of rational numbers and, for each prime p, there is a positive integer n for which $f(n)$ is not divisible by p. It was conjectured by Bouniakowsky (1857) that conversely, if these three conditions are satisfied, then $f(n)$ is prime for infinitely many positive integers n. Schinzel (1958) extended the conjecture to several polynomials and Bateman and Horn (1962) gave Schinzel's conjecture the following quantitative form.

Let $f_j(t)$ be a polynomial in t of degree $d_j \geq 1$, with integer coefficients and positive leading coefficient, which is irreducible over the field $\mathbb{Q}$ of rational numbers ($j = 1, \ldots, m$). Suppose also that the polynomials $f_1(t), \ldots, f_m(t)$ are distinct and that, for each prime p, there is a positive integer n for which the product $f_1(n) \cdots f_m(n)$ is not divisible by p. Bateman and Horn's conjecture states that, if $N(x)$ is the number of positive integers $n \leq x$ for which $f_1(n), \ldots, f_m(n)$ are all primes, then

$$N(x) \sim (d_1 \cdots d_m)^{-1} C(f_1, \ldots, f_m) \int_2^x dt / \log^m t,$$

where

$$C(f_1, \ldots, f_m) = \prod_p \{(1 - 1/p)^{-m}(1 - \omega(p)/p)\},$$

the product being taken over all primes p and $\omega(p)$ denoting the number of $u \in \mathbb{F}_p$ (the field of p elements) such that $f_1(u) \cdots f_m(u) = 0$. (The convergence of the infinite product when the primes are taken in their natural order follows from the prime ideal theorem.)

The twin prime formula is obtained by taking $m = 2$ and $f_1(t) = t$, $f_2(t) = t + 2$. By taking instead $f_1(t) = t$, $f_2(t) = 2t + 1$, the Bateman–Horn conjecture gives the same asymptotic formula $\pi_G(x) \sim L_2(x)$ for the number $\pi_G(x)$ of primes $p \leq x$ for which $2p + 1$ is also a prime ('Sophie Germain' primes). By taking $m = 1$ and $f_1(t) = t^2 + 1$ one obtains an asymptotic formula for the number of primes of the form $n^2 + 1$.

Bateman and Horn gave a heuristic derivation of their formula. However, the only case in which the formula has actually been proved is $m = 1$, $n_1 = 1$. This is the case of primes in an arithmetic progression which will be considered in the next chapter. When one considers the vast output of mathematical papers today compared with previous eras, it is salutary to recall that we still do not know as much about twin primes as Euclid knew about primes.

8 Further Remarks

The historical development of the prime number theorem is traced in Landau [33]. The original papers of Chebyshev are available in [56]. Pintz [48] has given a simple proof of Chebyshev's result that $\pi(x) = x/(A \log x - B + o(1))$ implies $A = B = 1$.

There is an English translation of Riemann's memoir in Edwards [20]. Complex variable proofs of the prime number theorem, with error term, are contained in the books of Ayoub [4], Ellison and Ellison [21], and Patterson [47]. For a simple complex variable proof without error term, due to Newman (1980), see Zagier [63].

A proof with error term by the Wiener–Ikehara method is given in Čižek [12]. Wiener's general Tauberian theorem is proved in Rudin [52]. For its algebraic interpretation, see the resumé of Fourier analysis in [13]. The development of Selberg's method is surveyed in Diamond [18]. An elementary proof of the prime number theorem which is quite different from that of Selberg and Erdős has been given by Daboussi [15].

A clear account of Stieltjes integrals is given in Widder [62]. However, we do not use Stieltjes integrals in any essential way, but only for the formal convenience of treating integration by parts and summation by parts in the same manner. Widder's book also contains the Wiener–Ikehara proof of the prime number theorem.

By a theorem of S. Bernstein (1928), proved in Widder's book and also in Mattner [38], the hypotheses of Proposition 7 can be stated without reference to the function $\phi(x)$. Bernstein's theorem says that a real-valued function $F(\sigma)$ can be represented in the form

$$F(\sigma) = \int_0^\infty e^{-\sigma x} d\phi(x),$$

where $\phi(x)$ is a nondecreasing function for $x \geq 0$ and the integral is convergent for every $\sigma > 1$, if and only if $F(\sigma)$ has derivatives of all orders and

$$(-1)^k F^{(k)}(\sigma) \geq 0 \quad \text{for every } \sigma > 1 \quad (k = 0, 1, 2, \ldots).$$

For the Poisson summation formula see, for example, Lasser [34] and Durán *et al.* [19]. There is a useful n-dimensional generalization, discussed more fully in §7 of Chapter XII, in which a sum over all points of a lattice is related to a sum over all points of the dual lattice. Further generalizations are mentioned in Chapter X.

More extended treatments of the gamma function are given in Andrews *et al.* [3] and Remmert [49].

More information about the Riemann zeta-function is given in the books of Patterson [47], Titchmarsh [57], and Karatsuba and Voronin [30]. For numerical data, see Rosser and Schoenfeld [50], van de Lune *et al.* [37] and Rumely [53].

For a proof that $\pi(x) - Li(x)$ changes sign infinitely often, see Diamond [17]. Estimates for values of x such that $\pi(x) > Li(x)$ are obtained by a technique due to Lehman [35]; for the most recent estimate, see Bays and Hudson [8].

For the pair correlation conjecture, see Montgomery [40], Goldston [24] and Odlyzko [45]. Random matrices are thoroughly discussed by Mehta [39]; for a nice introduction, see Tracy and Widom [58].

For Dedekind zeta functions see Stark [54], besides the books on algebraic number theory referred to in Chapter III. The prime ideal theorem is proved in Narkiewicz [44], for example. For consequences of the extended Riemann hypothesis, see Bach [5], Goldstein [23] and M.R. Murty [41]. Many other generalizations of the zeta function are discussed in the article on zeta functions in [22].

Function fields are treated in the books of Chevalley [11] and Deuring [16]. The lengthy review of Chevalley's book by Weil in *Bull. Amer. Math. Soc.* **57** (1951), 384–398, is useful but over-critical. Even if geometric methods are better adapted for algebraic varieties of higher dimension, the algebraic methods available for curves are essentially simpler. Moreover it was the close analogy with number fields that suggested the possibility of a Riemann hypothesis for function fields. For a proof of the latter, see Bombieri [9]. For the Weil conjectures, see Weil [61] and Katz [32].

Stichtenoth [55] gives a good account of the theory of function fields with special emphasis on its applications to coding theory. For these applications, see also Goppa [26], Tsfasman *et al.* [60], and Tsfasman and Vladut [59]. Curves with a given genus which have the maximal number of $\mathbb{F}_q$-points are discussed by Cossidente *et al.* [14].

For introductions to Ramanujan's tau-function, see V.K. Murty [42] and Rankin's article (pp. 245–268) in Andrews *et al.* [2]. For analogues of the prime number theorem in the theory of dynamical systems, see Katok and Hasselblatt [31] and Parry and Pollicott [46]. Hadamard's pioneering study of geodesics on a surface of negative curvature and his proof of the prime number theorem are both reproduced in [27].

The 'equivalence' of Proposition 12 with the prime number theorem is proved in Ayoub [4]. A proof that the Riemann hypothesis is equivalent to $M(x) = O(x^{\alpha})$ for every $\alpha > 1/2$ is contained in the book of Titchmarsh [57]. Good and Churchhouse's probabilistic conjectures appeared in [25]. For the central limit theorem and the law of the iterated logarithm see, for example, Adams [1], Kac [29], Bauer [7] and Loève [36].

Brun's theorem on twin primes is proved in Narkiewicz [43]. For numerical results, see Brent [10]. For conjectural asymptotic formulas, see Hardy and Littlewood [28] and Bateman and Horn [6]. There are several heuristic derivations of the twin prime formula, the most recent being Rubenstein [51]. It would be useful to try to analyse these heuristic derivations, so that the conclusion is seen as a consequence of precisely stated assumptions.

9 Selected References

[1] W.J. Adams, *The life and times of the central limit theorem*, Kaedmon, New York, 1974.

[2] G.E. Andrews, R.A. Askey, B.C. Berndt, K.G. Ramanathan and R.A. Rankin (ed.), *Ramanujan revisited*, Academic Press, London, 1988.

[3] G.E. Andrews, R. Askey and R. Roy, *Special functions*, Cambridge University Press, 1999.

[4] R. Ayoub, *An introduction to the analytic theory of numbers*, Math. Surveys no. 10, Amer. Math. Soc., Providence, 1963.
[5] E. Bach, Explicit bounds for primality testing and related problems, *Math. Comp.* **55** (1990), 353–380.
[6] P.T. Bateman and R.A. Horn, A heuristic asymptotic formula concerning the distribution of prime numbers, *Math. Comp.* **16** (1962), 363–367.
[7] H. Bauer, *Probability theory*, English transl. by R.B. Burckel, de Gruyter, Berlin, 1996.
[8] C. Bays and R.H. Hudson, A new bound for the smallest x with $\pi(x) > li(x)$, *Math. Comp.* **69** (1999), 1285–1296.
[9] E. Bombieri, Counting points on curves over finite fields (d'après S.A. Stepanov), *Séminaire Bourbaki vol. 1972/3, Exposés 418–435*, pp. 234–241, Lecture Notes in Mathematics **383** (1974), Springer-Verlag, Berlin.
[10] R.P. Brent, Irregularities in the distribution of primes and twin primes, *Math. Comp.* **29** (1975), 43–56.
[11] C. Chevalley, *Introduction to the theory of algebraic functions of one variable*, Math. Surveys no. 6, Amer. Math. Soc., New York, 1951.
[12] J. Čižek , On the proof of the prime number theorem, *Časopis Pěst. Mat.* **106** (1981), 395–401.
[13] W.A. Coppel, J.B. Fourier–On the occasion of his two hundredth birthday, *Amer. Math. Monthly* **76** (1969), 468–483.
[14] A. Cossidente, J.W.P. Hirschfeld, G. Korchmáros and F. Torres, On plane maximal curves, *Compositio Math.* **121** (2000), 163–181.
[15] H. Daboussi, Sur le théorème des nombres premiers, *C.R. Acad. Sci. Paris Sér. I* **298** (1984), 161–164.
[16] M. Deuring, *Lectures on the theory of algebraic functions of one variable*, Lecture Notes in Mathematics **314** (1973), Springer-Verlag, Berlin.
[17] H.G. Diamond, Changes of sign of $\pi(x) - li(x)$, *Enseign. Math.* **21** (1975), 1–14.
[18] H.G. Diamond, Elementary methods in the study of the distribution of prime numbers, *Bull. Amer. Math. Soc.* (*N.S.*) **7** (1982), 553–589.
[19] A.L. Durán, R. Estrada and R.P. Kanwal, Extensions of the Poisson summation formula, *J. Math. Anal. Appl.* **218** (1998), 581–606.
[20] H.M. Edwards, *Riemann's zeta function*, Academic Press, New York, 1974.
[21] W. Ellison and F. Ellison, *Prime numbers*, Wiley, New York, 1985.
[22] *Encyclopedic dictionary of mathematics* (ed. K. Ito), 2nd ed., Mathematical Society of Japan, MIT Press, Cambridge, Mass., 1987.
[23] L.J. Goldstein, Density questions in algebraic number theory, *Amer. Math. Monthly* **78** (1971), 342–351.
[24] D.A. Goldston, On the pair correlation conjecture for zeros of the Riemann zeta-function, *J. Reine Angew. Math.* **385** (1988), 24–40.
[25] I.J. Good and R.F. Churchhouse, The Riemann hypothesis and pseudorandom features of the Möbius sequence, *Math. Comp.* **22** (1968), 857–861.
[26] V.D. Goppa, Codes on algebraic curves, *Soviet Math. Dokl.* **24** (1981), 170–172.
[27] J. Hadamard, *Selecta*, Gauthier-Villars, Paris, 1935.
[28] G.H. Hardy and J.E. Littlewood, Some problems of partitio numerorum III, On the expression of a number as a sum of primes, *Acta Math.* **44** (1923), 1–70.
[29] M. Kac, *Statistical independence in probability, analysis and number theory*, Carus Mathematical Monograph **12**, Math. Assoc. of America, 1959.
[30] A.A. Karatsuba and S.M. Voronin, *The Riemann zeta-function*, English transl. by N. Koblitz, de Gruyter, Berlin, 1992.
[31] A. Katok and B. Hasselblatt, *Introduction to the modern theory of dynamical systems*, Cambridge University Press, Cambridge, 1995.

[32] N. Katz, An overview of Deligne's proof of the Riemann hypothesis for varieties over finite fields, *Mathematical developments arising from Hilbert problems*, Proc. Sympos. Pure Math. **28**, pp. 275–305, Amer. Math. Soc., Providence, 1976.
[33] E. Landau, *Handbuch der Lehre von der Verteilung der Primzahlen* (2 vols.), 2nd ed., Chelsea, New York, 1953.
[34] R. Lasser, *Introduction to Fourier series*, M. Dekker, New York, 1996.
[35] R.S. Lehman, On the difference $\pi(x) - li(x)$, *Acta Arith.* **11** (1966), 397–410.
[36] M. Loève, *Probability theory*, 4th ed. in 2 vols., Springer-Verlag, New York, 1978.
[37] J. van de Lune *et al*., On the zeros of the Riemann zeta function in the critical strip IV, *Math. Comp.* **46** (1986), 667–681.
[38] L. Mattner, Bernstein's theorem, inversion formula of Post and Widder, and the uniqueness theorem for Laplace transforms, *Exposition. Math.* **11** (1993), 137–140.
[39] M.L. Mehta, *Random matrices*, 2nd ed., Academic Press, New York, 1991.
[40] H.L. Montgomery, The pair correlation of zeros of the zeta function, *Proc. Sympos. Pure Math.* **24**, pp. 181–193, Amer. Math. Soc., Providence, 1973.
[41] M. R. Murty, Artin's conjecture for primitive roots, *Math. Intelligencer* **10** (1988), no. 4, 59–67.
[42] V.K. Murty, Ramanujan and Harish-Chandra, *Math. Intelligencer* **15** (1993), no.2, 33–39.
[43] W. Narkiewicz, *Number theory*, World Scientific, Singapore, 1983.
[44] W. Narkiewicz, *Elementary and analytic theory of algebraic numbers*, 2nd ed., Springer-Verlag, Berlin, 1990.
[45] A.M. Odlyzko, On the distribution of spacings between zeros of the zeta function, *Math. Comp.* **48** (1987), 273–308.
[46] W. Parry and M. Pollicott, An analogue of the prime number theorem for closed orbits of Axiom A flows, *Ann. of Math.* **118** (1983), 573–591.
[47] S.J. Patterson, *An introduction to the theory of the Riemann zeta-function*, Cambridge University Press, Cambridge, 1988.
[48] J. Pintz, On Legendre's prime number formula, *Amer. Math. Monthly* **87** (1980), 733–735.
[49] R. Remmert, *Classical topics in complex function theory*, English transl. by L. Kay, Springer-Verlag, New York, 1998.
[50] J.B. Rosser and L. Schoenfeld, Approximate formulas for some functions of prime numbers, *Illinois J. Math.* **6** (1962), 64–94.
[51] M. Rubinstein, A simple heuristic proof of Hardy and Littlewood's conjecture B, *Amer. Math. Monthly* **100** (1993), 456–460.
[52] W. Rudin, *Functional analysis*, McGraw-Hill, New York, 1973.
[53] R. Rumely, Numerical computations concerning the ERH, *Math. Comp.* **61** (1993), 415–440.
[54] H.M. Stark, The analytic theory of algebraic numbers, *Bull. Amer. Math. Soc.* **81** (1975), 961–972.
[55] H. Stichtenoth, *Algebraic function fields and codes*, Springer-Verlag, Berlin, 1993.
[56] P.L. Tchebychef, *Oeuvres* (2 vols.), reprinted Chelsea, New York, 1962.
[57] E.C. Titchmarsh, *The theory of the Riemann zeta-function*, 2nd ed. revised by D.R. Heath-Brown, Clarendon Press, Oxford, 1986.
[58] C.A. Tracy and H. Widom, Introduction to random matrices, *Geometric and quantum aspects of integrable systems* (ed. G.F. Helminck), pp. 103–130, Lecture Notes in Physics **424**, Springer-Verlag, Berlin, 1993.
[59] M.A. Tsfasman and S.G. Vladut, *Algebraic-geometric codes*, Kluwer, Dordrecht, 1991.
[60] M.A. Tsfasman, S.G. Vladut and Th. Zink, Modular curves, Shimura curves, and Goppa codes, *Math. Nachr.* **109** (1982), 21–28.
[61] A. Weil, Number of solutions of equations in finite fields, *Bull. Amer. Math. Soc.* **55** (1949), 497–508.

[62] D.V. Widder, *The Laplace transform*, Princeton University Press, Princeton, 1941.
[63] D. Zagier, Newman's short proof of the prime number theorem, *Amer. Math. Monthly* **104** (1997), 705–708.

Additional References

J.B. Conrey, The Riemann hypothesis, *Notices Amer. Math. Soc.* **50** (2003), 341–353.
N.M. Katz and P. Sarnak, Zeroes of zeta functions and symmetry, *Bull. Amer. Math. Soc.* (*N.S.*) **36** (1999), 1–26.

X

A Character Study

1 Primes in Arithmetic Progressions

Let a and m be integers with $1 \leq a < m$. If a and m have a common divisor $d > 1$, then no term after the first of the arithmetic progression

$$a, a+m, a+2m, \ldots \qquad (*)$$

is a prime. Legendre (1788) conjectured, and later (1808) attempted a proof, that *if a and m are relatively prime, then the arithmetic progression $(*)$ contains infinitely many primes*.

If $a_1, \ldots, a_h$ are the positive integers less than m and relatively prime to m, and if $\pi_j(x)$ denotes the number of primes $\leq x$ in the arithmetic progression

$$a_j, a_j+m, a_j+2m, \ldots,$$

then Legendre's conjecture can be stated in the form

$$\pi_j(x) \to \infty \quad \text{as } x \to \infty \quad (j = 1, \ldots, h).$$

Legendre (1830) subsequently conjectured, and again gave a faulty proof, that

$$\pi_j(x)/\pi_k(x) \to 1 \quad \text{as } x \to \infty \quad \text{for all } j, k.$$

Since the total number $\pi(x)$ of primes $\leq x$ satisfies

$$\pi(x) = \pi_1(x) + \cdots + \pi_h(x) + c,$$

where c is the number of different primes dividing m, Legendre's second conjecture is equivalent to

$$\pi_j(x)/\pi(x) \to 1/h \quad \text{as } x \to \infty \quad (j = 1, \ldots, h).$$

Here $h = \varphi(m)$ is the number of positive integers less than m and relatively prime to m. If one assumes the truth of the prime number theorem, then the second conjecture is

W.A. Coppel, *Number Theory: An Introduction to Mathematics*, Universitext,
DOI: 10.1007/978-0-387-89486-7_10, © Springer Science + Business Media, LLC 2009

also equivalent to

$$\pi_j(x) \sim x/\varphi(m)\log x \quad (j = 1, \ldots, \varphi(m)).$$

The validity of the second conjecture in this form is known as the *prime number theorem for arithmetic progressions*.

Legendre's first conjecture was proved by Dirichlet (1837) in an outstanding paper which combined number theory, algebra and analysis. His algebraic innovation was the use of *characters* to isolate the primes belonging to a particular residue class mod m. Legendre's second conjecture, which implies the first, was proved by de la Vallée Poussin (1896), again using characters, at the same time that he proved the ordinary prime number theorem.

Selberg (1949), (1950) has given proofs of both conjectures which avoid the use of complex analysis, but they are not very illuminating. The prime number theorem for arithmetic progressions will be proved here by an extension of the method used in the previous chapter to prove the ordinary prime number theorem.

For any integer a, with $1 \le a < m$ and $(a, m) = 1$, let

$$\pi(x; m, a) = \sum_{p \le x,\, p \equiv a \bmod m} 1.$$

Also, generalizing the definition of Chebyshev's functions in the previous chapter, put

$$\theta(x; m, a) = \sum_{p \le x,\, p \equiv a \bmod m} \log p, \quad \psi(x; m, a) = \sum_{n \le x,\, n \equiv a \bmod m} \Lambda(n).$$

Exactly as in the last chapter, we can show that the prime number theorem for arithmetic progressions,

$$\pi(x; m, a) \sim x/\varphi(m)\log x \quad \text{as } x \to \infty,$$

is equivalent to

$$\psi(x; m, a) \sim x/\varphi(m) \quad \text{as } x \to \infty.$$

It is in this form that the theorem will be proved.

2 Characters of Finite Abelian Groups

Let G be an abelian group with identity element e. A *character* of G is a function $\chi : G \to \mathbb{C}$ such that

(i) $\chi(ab) = \chi(a)\chi(b)$ for all $a, b \in G$,
(ii) $\chi(c) \ne 0$ for some $c \in G$.

Since $\chi(c) = \chi(ca^{-1})\chi(a)$, by (i), it follows from (ii) that $\chi(a) \ne 0$ for every $a \in G$. (Thus χ is a *homomorphism* of G into the multiplicative group $\mathbb{C}^\times$ of nonzero complex numbers.) Moreover, since $\chi(a) = \chi(a)\chi(e)$, we must have $\chi(e) = 1$. Since $\chi(a)\chi(a^{-1}) = \chi(e)$, it follows that $\chi(a^{-1}) = \chi(a)^{-1}$.

The function $\chi_1 : G \to \mathbb{C}$ defined by $\chi_1(a) = 1$ for every $a \in G$ is obviously a character of G, the *trivial character* (also called the *principal* character!). Moreover, for any character χ of G, the function $\chi^{-1} : G \to \mathbb{C}$ defined by $\chi^{-1}(a) = \chi(a)^{-1}$ is also a character of G. Furthermore, if χ' and χ'' are characters of G, then the function $\chi'\chi'' : G \to \mathbb{C}$ defined by $\chi'\chi''(a) = \chi'(a)\chi''(a)$ is a character of G. Since

$$\chi_1\chi = \chi, \quad \chi'\chi'' = \chi''\chi', \quad \chi(\chi'\chi'') = (\chi\chi')\chi'',$$

it follows that the set $\hat{G}$ of all characters of G is itself an abelian group, the *dual group* of G, with the trivial character as identity element.

Suppose now that the group G is finite, of order g say. Then $\chi(a)$ is a g-th root of unity for every $a \in G$, since $a^g = e$ and hence

$$\chi(a)^g = \chi(a^g) = \chi(e) = 1.$$

It follows that $|\chi(a)| = 1$ and $\chi^{-1}(a) = \overline{\chi(a)}$. Thus we will sometimes write $\bar{\chi}$ instead of χ^{-1}.

Proposition 1 *The dual group $\hat{G}$ of a finite abelian group G is a finite abelian group of the same order. Moreover, if $a \in G$ and $a \neq e$, then $\chi(a) \neq 1$ for some $\chi \in \hat{G}$.*

Proof Let g denote the order of G. Suppose first that G is a cyclic group, generated by the element c. Then any character χ of G is uniquely determined by the value $\chi(c)$, which is a g-th root of unity. Conversely if $\omega_j = e^{2\pi ij/g}$ $(0 \leq j < g)$ is a g-th root of unity, then the functions $\chi^{(j)} : G \to \mathbb{C}$ defined by $\chi^{(j)}(c^k) = \omega_j^k$ are distinct characters of G and $\chi^{(1)}(c^k) \neq 1$ for $1 \leq k < g$. It follows that the proposition is true when G is cyclic. The general case can be reduced to this by using the fact (see §4 of Chapter III) that any finite abelian group is a direct product of cyclic groups. However, it can also be treated directly in the following way.

We use induction on g and suppose that G is not cyclic. Let H be a maximal proper subgroup of G and let h be the order of H. Let $a \in G \backslash H$ and let r be the least positive integer such that $b = a^r \in H$. Since G is generated by H and a, and $a^n \in H$ if and only if r divides n, each $x \in G$ can be uniquely expressed in the form

$$x = a^k y,$$

where $y \in H$ and $0 \leq k < r$. Hence $g = rh$.

If χ is any character of G, its restriction to H is a character ψ of H. Moreover χ is uniquely determined by ψ and the value $\chi(a)$, since

$$\chi(a^k y) = \chi(a)^k \psi(y).$$

Since $\chi(a)^r = \psi(b)$ is a root of unity, $\omega = \chi(a)$ is a root of unity such that $\omega^r = \psi(b)$.

Conversely, it is easily verified that, for each character ψ of H and for each of the r roots of unity ω such that $\omega^r = \psi(b)$, the function $\chi : G \to \mathbb{C}$ defined by $\chi(a^k y) = \omega^k \psi(y)$ is a character of G. Since H has exactly h characters by the induction hypothesis, it follows that G has exactly $rh = g$ characters. It remains to show that if $a^k y \neq e$, then $\chi(a^k y) \neq 1$ for some χ. But if $\omega^k \psi(y) = 1$ for all ω, then $k = 0$; hence $y \neq e$ and $\chi(y) = \psi(y) \neq 1$ for some ψ, by the induction hypothesis. □

Proposition 2 *Let G be a finite abelian group of order g and $\hat{G}$ its dual group. Then*

(i)

$$\sum_{a\in G}\chi(a)=\begin{cases}g & \text{if } \chi=\chi_1,\\ 0 & \text{if } \chi\neq\chi_1.\end{cases}$$

(ii)

$$\sum_{\chi\in\hat{G}}\chi(a)=\begin{cases}g & \text{if } a=e,\\ 0 & \text{if } a\neq e.\end{cases}$$

Proof Put

$$S=\sum_{a\in G}\chi(a).$$

Since it is obvious that $S=g$ if $\chi=\chi_1$, we assume $\chi\neq\chi_1$. Then $\chi(b)\neq 1$ for some $b\in G$. Since ab runs through all elements of G at the same time as a,

$$\chi(b)S=\sum_{a\in G}\chi(a)\chi(b)=\sum_{a\in G}\chi(ab)=S.$$

Since $\chi(b)\neq 1$, it follows that $S=0$.

Now put

$$T=\sum_{\chi\in\hat{G}}\chi(a).$$

Evidently $T=g$ if $a=e$ since, by Proposition 1, $\hat{G}$ also has order g. Thus we now assume $a\neq e$. By Proposition 1 also, for some $\psi\in\hat{G}$ we have $\psi(a)\neq 1$. Since $\chi\psi$ runs through all elements of $\hat{G}$ at the same time as χ,

$$\psi(a)T=\sum_{\chi\in\hat{G}}\chi(a)\psi(a)=\sum_{\chi\in\hat{G}}\chi\psi(a)=T.$$

Since $\psi(a)\neq 1$, it follows that $T=0$. □

Since the product of two characters is again a character, and since $\bar{\psi}$ is the inverse of the character ψ, Proposition 2(i) can be stated in the apparently more general form

(i)′
$$\sum_{a\in G}\chi(a)\bar{\psi}(a)=\begin{cases}g & \text{if } \chi=\psi,\\ 0 & \text{if } \chi\neq\psi.\end{cases}$$

Similarly, since $\bar{\chi}(b)=\chi(b^{-1})$, Proposition 2(ii) can be stated in the form

(ii)′
$$\sum_{\chi\in\hat{G}}\chi(a)\bar{\chi}(b)=\begin{cases}g & \text{if } a=b,\\ 0 & \text{if } a\neq b.\end{cases}$$

The relations (i)′ and (ii)′ are known as the *orthogonality relations*, for the characters and elements respectively, of a finite abelian group.

3 Proof of the Prime Number Theorem for Arithmetic Progressions

The finite abelian group in which we are interested is the multiplicative group $\mathbb{Z}_{(m)}^{\times}$ of integers relatively prime to m, where $m > 1$ will be fixed from now on. The group $G_m = \mathbb{Z}_{(m)}^{\times}$ has order $\varphi(m)$, where $\varphi(m)$ denotes as usual the number of positive integers less than m and relatively prime to m.

A *Dirichlet character* mod m is defined to be a function $\chi : \mathbb{Z} \to \mathbb{C}$ with the properties

(i) $\chi(ab) = \chi(a)\chi(b)$ for all $a, b \in \mathbb{Z}$,
(ii) $\chi(a) = \chi(b)$ if $a \equiv b \bmod m$,
(iii) $\chi(a) \neq 0$ if and only if $(a, m) = 1$.

Any character χ of G_m can be extended to a Dirichlet character mod m by putting $\chi(a) = 0$ if $a \in \mathbb{Z}$ and $(a, m) \neq 1$. Conversely, on account of (ii), any Dirichlet character mod m uniquely determines a character of G_m.

To illustrate the definition, here are some examples of Dirichlet characters. In each case we set $\chi(a) = 0$ if $(a, m) \neq 1$.

(I) $m = p$ is an odd prime and $\chi(a) = (a/p)$ if $p \nmid a$, where (a/p) is the Legendre symbol;
(II) $m = 4$ and $\chi(a) = 1$ or -1 according as $a \equiv 1$ or $-1 \bmod 4$;
(III) $m = 8$ and $\chi(a) = 1$ or -1 according as $a \equiv \pm 1$ or $\pm 3 \bmod 8$.

We now return to the general case. By the results of the previous section we have

$$\sum_{n=1}^{m} \chi(n) \equiv \begin{cases} \varphi(m) & \text{if } \chi = \chi_1, \\ 0 & \text{if } \chi \neq \chi_1, \end{cases}$$

and

$$\sum_{\chi} \chi(a) = \begin{cases} \varphi(m) & \text{if } a \equiv 1 \bmod m, \\ 0 & \text{otherwise,} \end{cases}$$

where χ runs through all Dirichlet characters mod m. Furthermore

$$\sum_{n=1}^{m} \chi(n)\bar{\psi}(n) = \begin{cases} \varphi(m) & \text{if } \chi = \psi, \\ 0 & \text{if } \chi \neq \psi, \end{cases}$$

and

$$\sum_{\chi} \chi(a)\bar{\chi}(b) = \begin{cases} \varphi(m) & \text{if } (a, m) = 1 \text{ and } a \equiv b \bmod m, \\ 0 & \text{otherwise.} \end{cases}$$

Lemma 3 *If $\chi \neq \chi_1$ is a Dirichlet character* mod m *then, for any positive integer N,*

$$\left| \sum_{n=1}^{N} \chi(n) \right| \leq \varphi(m)/2.$$

Proof Any positive integer N can be written in the form $N = qm + r$, where $q \geq 0$ and $1 \leq r \leq m$. Since $\chi(a) = \chi(b)$ if $a \equiv b \bmod m$, we have

$$\sum_{n=1}^{N} \chi(n) = \left(\sum_{n=1}^{m} + \sum_{n=m+1}^{2m} + \cdots + \sum_{n=(q-1)m+1}^{qm} \right) \chi(n) + \sum_{n=qm+1}^{qm+r} \chi(n)$$
$$= q \sum_{n=1}^{m} \chi(n) + \sum_{n=1}^{r} \chi(n).$$

But $\sum_{n=1}^{m} \chi(n) = 0$, since $\chi \neq \chi_1$. Hence

$$\sum_{n=1}^{N} \chi(n) = \sum_{n=1}^{r} \chi(n) = -\sum_{n=r+1}^{m} \chi(n).$$

Since $|\chi(n)| = 1$ or 0 according as $(n, m) = 1$ or $(n, m) \neq 1$, and since $\varphi(m)$ is the number of positive integers $n \leq m$ such that $(n, m) = 1$, the result follows. □

With each Dirichlet character χ, there is associated a *Dirichlet L-function*

$$L(s, \chi) = \sum_{n=1}^{\infty} \chi(n)/n^s.$$

Since $|\chi(n)| \leq 1$ for all n, the series is absolutely convergent for $\sigma := \mathscr{R}s > 1$. We are going to show that if $\chi \neq \chi_1$, then the series is also convergent for $\sigma > 0$. (It does not converge if $\sigma \leq 0$, since then $|\chi(n)/n^s| \geq 1$ for infinitely many n.)

Put

$$H(x) = \sum_{n \leq x} \chi(n).$$

Then

$$\sum_{n \leq x} \chi(n) n^{-s} = \int_{1-}^{x+} t^{-s} \, dH(t)$$
$$= H(x) x^{-s} + s \int_{1}^{x} H(t) t^{-s-1} dt.$$

Since $H(x)$ is bounded, by Lemma 3, on letting $x \to \infty$ we obtain

$$L(s, \chi) = s \int_{1}^{\infty} H(t) t^{-s-1} dt \quad \text{for } \sigma > 0.$$

Moreover the integral on the right is uniformly convergent in any half-plane $\sigma \geq \delta$, where $\delta > 0$, and hence $L(s, \chi)$ is a holomorphic function for $\sigma > 0$.

The following discussion of Dirichlet L-functions and the prime number theorem for arithmetic progressions runs parallel to that of the Riemann ζ-function and the ordinary prime number theorem in the previous chapter. Consequently we will be more brief.

Proposition 4 *$L(s,\chi) = \Pi_p(1-\chi(p)p^{-s})^{-1}$ for $\sigma > 1$, where the product is taken over all primes p.*

Proof The property $\chi(ab) = \chi(a)\chi(b)$ for all $a, b \in \mathbb{N}$ enables the proof of Euler's product formula for $\zeta(s)$ to be carried over to the present case. For $\sigma > 0$ we have

$$(1-\chi(p)p^{-s})^{-1} = 1 + \chi(p)p^{-s} + \chi(p^2)p^{-2s} + \chi(p^3)p^{-3s} + \cdots$$

and hence for $\sigma > 1$

$$\prod_{p\le x}(1-\chi(p)p^{-s})^{-1} = \sum_{n\le N_x}\chi(n)n^{-s},$$

where N_x is the set of all positive integers whose prime factors are all $\le x$. Letting $x \to \infty$, we obtain the result. □

It follows at once that

$$L(s,\chi_1) = \zeta(s)\prod_{p|m}(1-p^{-s})$$

and that, for any Dirichlet character χ, $L(s,\chi) \neq 0$ for $\sigma > 1$.

Proposition 5 *$-L'(s,\chi)/L(s,\chi) = \sum_{n=1}^{\infty}\chi(n)\Lambda(n)/n^s$ for $\sigma > 1$.*

Proof The series $\omega(s,\chi) = \sum_{n=1}^{\infty}\chi(n)\Lambda(n)n^{-s}$ converges absolutely and uniformly in any half-plane $\sigma \ge 1+\varepsilon$, where $\varepsilon > 0$. Moreover, as in the proof of Proposition IX.6,

$$L(s,\chi)\omega(s,\chi) = \sum_{j=1}^{\infty}\chi(j)j^{-s}\sum_{k=1}^{\infty}\chi(k)\Lambda(k)k^{-s} = \sum_{n=1}^{\infty}n^{-s}\sum_{jk=n}\chi(j)\chi(k)\Lambda(k)$$

$$= \sum_{n=1}^{\infty}n^{-s}\chi(n)\sum_{d|n}\Lambda(d) = \sum_{n=1}^{\infty}n^{-s}\chi(n)\log n = -L'(s,\chi). \qquad \square$$

As in the proof of Proposition IX.6, we can also prove directly that $L(s,\chi) \neq 0$ for $\sigma > 1$, and thus make the proof of the prime number theorem for arithmetic progressions independent of Proposition 4.

The following general result, due to Landau (1905), considerably simplifies the subsequent argument (and has other applications).

Proposition 6 *Let $\phi(x)$ be a nondecreasing function for $x \ge 0$ such that the integral*

$$f(s) = \int_0^{\infty} e^{-sx}d\phi(x) \qquad (\dagger)$$

is convergent for $\Re s > \beta$. Thus f is holomorphic in this half-plane. If the definition of f can be extended so that it is holomorphic on the real segment $(\alpha,\beta]$, then the integral in (†) is convergent also for $\Re s > \alpha$. Thus f is actually holomorphic, and (†) holds, in this larger half-plane.

Proof Since f is holomorphic at β, we can choose $\delta > 0$ so that f is holomorphic in the disc $|s - (\beta + \delta)| < 2\delta$. Thus its Taylor series converges in this disc. But for $\mathscr{R}s > \beta$ the n-th derivative of f is given by

$$f^{(n)}(s) = (-1)^n \int_0^\infty e^{-sx} x^n d\phi(x).$$

Hence, for any σ such that $\beta - \delta < \sigma < \beta + \delta$,

$$\begin{aligned} f(\sigma) &= \sum_{n=0}^{\infty} (\sigma - \beta - \delta)^n f^{(n)}(\beta + \delta)/n! \\ &= \sum_{n=0}^{\infty} (\sigma - \beta - \delta)^n (-1)^n \int_0^\infty e^{-(\beta+\delta)x} x^n d\phi(x)/n! \\ &= \sum_{n=0}^{\infty} \int_0^\infty e^{-(\beta+\delta)x} (\beta + \delta - \sigma)^n x^n / n!\, d\phi(x). \end{aligned}$$

Since the integrands are non-negative, we can interchange the orders of summation and integration, obtaining

$$\begin{aligned} f(\sigma) &= \int_0^\infty e^{-(\beta+\delta)x} \sum_{n=0}^{\infty} (\beta + \delta - \sigma)^n x^n / n!\, d\phi(x) \\ &= \int_0^\infty e^{-(\beta+\delta)x} e^{(\beta+\delta-\sigma)x} d\phi(x) \\ &= \int_0^\infty e^{-\sigma x} d\phi(x). \end{aligned}$$

Thus the integral in (†) converges for real $s > \beta - \delta$.

Let γ be the greatest lower bound of all real $s \in (\alpha, \beta)$ for which the integral in (†) converges. Then the integral in (†) is also convergent for $\mathscr{R}s > \gamma$ and defines there a holomorphic function. Since this holomorphic function coincides with $f(s)$ for $\mathscr{R}s > \beta$, it follows that (†) holds for $\mathscr{R}s > \gamma$. Moreover $\gamma = \alpha$, since if $\gamma > \alpha$ we could replace β by γ in the preceding argument and thus obtain a contradiction to the definition of γ. □

The punch-line is the following proposition:

Proposition 7 *$L(1 + it, \chi) \neq 0$ for every real t and every $\chi \neq \chi_1$.*

Proof Assume on the contrary that $L(1 + i\alpha, \chi) = 0$ for some real α and some $\chi \neq \chi_1$. Then also $L(1 - i\alpha, \bar{\chi}) = 0$. If we put

$$f(s) = \zeta^2(s) L(s + i\alpha, \chi) L(s - i\alpha, \bar{\chi}),$$

then f is holomorphic and nonzero for $\sigma > 1$. Furthermore f is holomorphic on the real segment $[1/2, 1]$, since the double pole of $\zeta^2(s)$ at $s = 1$ is cancelled by the zeros of the other two factors. By logarithmic differentiation we obtain, for $\sigma > 1$,

$$\begin{aligned}
&-f'(s)/f(s)\\
&\quad = -2\zeta'(s)/\zeta(s) - L'(s+i\alpha,\chi)/L(s+i\alpha,\chi) - L'(s-i\alpha,\bar{\chi})/L(s-i\alpha,\bar{\chi})\\
&\quad = 2\sum_{n=1}^{\infty} \Lambda(n)n^{-s} + \sum_{n=1}^{\infty} \chi(n)\Lambda(n)n^{-s-i\alpha} + \sum_{n=1}^{\infty} \bar{\chi}(n)\Lambda(n)n^{-s+i\alpha}\\
&\quad = \sum_{n=2}^{\infty} c_n n^{-s},
\end{aligned}$$

where

$$c_n = \{2 + \chi(n)n^{-i\alpha} + \bar{\chi}(n)n^{i\alpha}\}\Lambda(n) = 2\{1 + \mathscr{R}(\chi(n)n^{-i\alpha})\}\Lambda(n).$$

Since $|\chi(n)| \le 1$ and $|n^{-i\alpha}| = 1$, it follows that $c_n \ge 0$ for all $n \ge 2$. If we put

$$g(s) = \sum_{n=2}^{\infty} c_n n^{-s}/\log n,$$

then $g'(s) = f'(s)/f(s)$ for $\sigma > 1$ and so the derivative of $e^{-g(s)}f(s)$ is

$$\{f'(s) - g'(s)f(s)\}e^{-g(s)} = 0.$$

Thus $f(s) = Ce^{g(s)}$, where C is a constant. In fact $C = 1$, since $g(\sigma) \to 0$ and $f(\sigma) \to 1$ as $\sigma \to +\infty$. Since $g(s)$ is the sum of an absolutely convergent Dirichlet series with nonnegative coefficients, so also are the powers $g^k(s)$ $(k = 2, 3, \ldots)$. Hence also

$$f(s) = e^{g(s)} = 1 + g(s) + g^2(s)/2! + \cdots = \sum_{n=1}^{\infty} a_n n^{-s} \quad \text{for } \sigma > 1,$$

where $a_n \ge 0$ for every n. It follows from Proposition 6 that the series $\sum_{n=1}^{\infty} a_n n^{-\sigma}$ must actually converge with sum $f(\sigma)$ for $\sigma \ge 1/2$. We will show that this leads to a contradiction.

Take $n = p^2$, where p is a prime. Then, by the manner of its formation,

$$\begin{aligned}
a_n &\ge c_n/\log n + c_p^2/2\log^2 p\\
&= \{2 + \chi(p)^2 p^{-2i\alpha} + \bar{\chi}(p)^2 p^{2i\alpha}\}/2 + \{2 + \chi(p)p^{-i\alpha} + \bar{\chi}(p)p^{i\alpha}\}^2/2\\
&= 2 - \chi(p)\bar{\chi}(p) + \{1 + \chi(p)p^{-i\alpha} + \bar{\chi}(p)p^{i\alpha}\}^2 \ge 1,
\end{aligned}$$

since $|\chi(p)| \le 1$. Hence

$$f(1/2) = \sum_{n=1}^{\infty} a_n/n^{1/2} \ge \sum_{n=p^2} a_n/n^{1/2} \ge \sum_{p} 1/p.$$

Since $\sum_p 1/p$ diverges, this is a contradiction. □

Proposition 8 $\sum_{n\le x}\chi_1(n)\Lambda(n)\sim x$, $\sum_{n\le x}\chi(n)\Lambda(n)=o(x)$ *if* $\chi\neq\chi_1$.

Proof For any Dirichlet character χ, put

$$g(s)=-\zeta'(s)/\zeta(s)-L'(s,\chi)/2L(s,\chi)-L'(s,\bar\chi)/2L(s,\bar\chi),$$
$$h(s)=-\zeta'(s)/\zeta(s)-L'(s,\chi)/2iL(s,\chi)+L'(s,\bar\chi)/2iL(s,\bar\chi).$$

For $\sigma=\mathscr{R}s>1$ we have

$$g(s)=\sum_{n=1}^{\infty}\{1+\mathscr{R}\chi(n)\}\Lambda(n)n^{-s},$$
$$h(s)=\sum_{n=1}^{\infty}\{1+\mathscr{I}\chi(n)\}\Lambda(n)n^{-s}.$$

If $\chi\neq\chi_1$ then, by Proposition 7, $g(s)-1/(s-1)$ and $h(s)-1/(s-1)$ are holomorphic for $\mathscr{R}s\ge 1$. Since the coefficients of the Dirichlet series for $g(s)$ and $h(s)$ are nonnegative, it follows from Ikehara's theorem (Theorem IX.9) that

$$\sum_{n\le x}\{1+\mathscr{R}\chi(n)\}\Lambda(n)\sim x,$$
$$\sum_{n\le x}\{1+\mathscr{I}\chi(n)\}\Lambda(n)\sim x.$$

On the other hand, if $\chi=\chi_1$ then $g(s)-2/(s-1)$ and $h(s)-1/(s-1)$ are holomorphic for $\mathscr{R}s\ge 1$, from which we obtain in the same way

$$\sum_{n\le x}\{1+\chi_1(n)\}\Lambda(n)\sim 2x,$$
$$\sum_{n\le x}\Lambda(n)\sim x.$$

The result follows. □

The prime number theorem for arithmetic progressions can now be deduced immediately. For, by the orthogonality relations and Proposition 8, if $1\le a<m$ and $(a,m)=1$, then

$$\begin{aligned}\psi(x;m,a)&=\sum_{n\le x,\,n\equiv a \bmod m}\Lambda(n)\\&=\sum_{\chi}\bar\chi(a)\sum_{n\le x}\chi(n)\Lambda(n)/\varphi(m)\\&\sim x/\varphi(m).\end{aligned}$$

It is also possible to obtain error bounds in the prime number theorem for arithmetic progressions of the same type as those in the ordinary prime number theorem. For example, it may be shown that for each $\alpha>0$,

$$\psi(x; m, a) = x/\varphi(m) + O(x/\log^{\alpha} x),$$
$$\pi(x; m, a) = Li(x)/\varphi(m) + O(x/\log^{\alpha} x),$$

where the constants implied by the O-symbols depend on α, but not on m or a.

In the same manner as for the Riemann zeta-function $\zeta(s)$ it may be shown that the Dirichlet L-function $L(s, \chi)$ satisfies a functional equation, provided χ is a primitive character. (Here a Dirichlet character $\chi \bmod m$ is *primitive* if for each proper divisor d of m there exists an integer $a \equiv 1 \bmod d$ with $(a, m) = 1$ and $\chi(a) \neq 1$.) Explicitly, if χ is a primitive character mod m and if one puts

$$\Lambda(s, \chi) = (m/\pi)^{s/2}\Gamma((s + \delta)/2)L(s, \chi),$$

where $\delta = 0$ or 1 according as $\chi(-1) = 1$ or -1, then

$$\Lambda(1 - s, \bar{\chi}) = \varepsilon_{\chi}\Lambda(s, \chi),$$

where

$$\varepsilon_{\chi} = i^{-\delta}m^{-1/2}\sum_{k=1}^{m}\bar{\chi}(k)e^{2\pi ik/m}.$$

It follows from the functional equation that $|\varepsilon_{\chi}| = 1$. Indeed, by taking complex conjugates we obtain, for real s,

$$\Lambda(1 - s, \chi) = \bar{\varepsilon}_{\chi}\Lambda(s, \bar{\chi})$$

and hence, on replacing s by $1 - s$,

$$\Lambda(s, \chi) = \bar{\varepsilon}_{\chi}\Lambda(1 - s, \bar{\chi}) = \varepsilon_{\chi}\bar{\varepsilon}_{\chi}\Lambda(s, \chi).$$

The extended Riemann hypothesis implies that no Dirichlet L-function $L(s, \chi)$ has a zero in the half-plane $\mathscr{R}s > 1/2$, since $f(s) = \prod_{\chi} L(s, \chi)$ is the Dedekind zeta-function of the algebraic number field $K = \mathbb{Q}(e^{2\pi i/m})$. Hence it may be shown that if the extended Riemann hypothesis holds, then

$$\psi(x; m, a) = x/\varphi(m) + O(x^{1/2}\log^2 x)$$

and

$$\pi(x; m, a) = Li(x)/\varphi(m) + O(x^{1/2}\log x),$$

where the constants implied by the O-symbols are independent of m and a. Assuming the extended Riemann hypothesis, Bach and Sorenson (1996) have shown that, for any a, m with $1 \leq a < m$ and $(a, m) = 1$, the least prime $p \equiv a \bmod m$ satisfies $p < 2(m \log m)^2$.

Without any hypothesis, Linnik (1944) proved that there exists an absolute constant L such that the least prime in any arithmetic progression $a, a+m, a+2m, \ldots$, where $1 \leq a < m$ and $(a, m) = 1$, does not exceed m^L if m is sufficiently large. Heath-Brown (1992) has shown that one can take any $L > 11/2$.

4 Representations of Arbitrary Finite Groups

The problem of extending the character theory of finite abelian groups to arbitrary finite groups was proposed by Dedekind and solved by Frobenius (1896). Simplifications were afterwards found by Frobenius himself, Burnside and Schur (1905). We will follow Schur's treatment, which is distinguished by its simplicity. It turns out that for nonabelian groups the concept of 'representation' is more fundamental than that of 'character'.

A *representation* of a group G is a mapping ρ of G into the set of all linear transformations of a finite-dimensional vector space V over the field $\mathbb{C}$ of complex numbers which preserves products, i.e.

$$\rho(st) = \rho(s)\rho(t) \quad \text{for all } s, t \in G, \tag{1}$$

and maps the identity element of G into the identity transformation of V : $\rho(e) = I$. The dimension of the vector space V is called the *degree* of the representation (although 'dimension' would be more natural).

It follows at once from (1) that

$$\rho(s)\rho(s^{-1}) = \rho(s^{-1})\rho(s) = I.$$

Thus, for every $s \in G$, $\rho(s)$ is an invertible linear transformation of V and $\rho(s^{-1}) = \rho(s)^{-1}$. (Hence a representation of G is a *homomorphism* of G into the group $GL(V)$ of all invertible linear transformations of V.)

Any group has a *trivial representation* of degree 1 in which every element of the group is mapped into the scalar 1.

Also, with any group G of finite order g a representation of degree g may be defined in the following way. Let $s_1, \ldots, s_g$ be an enumeration of the elements of G and let $e_1, \ldots, e_g$ be a basis for a g-dimensional vector space V over $\mathbb{C}$. We define a linear transformation $A(s_i)$ of V by its action on the basis elements:

$$A(s_i)e_j = e_k \quad \text{if } s_i s_j = s_k.$$

Then, for all $s, t \in G$,

$$A(s^{-1})A(s) = I, \quad A(st) = A(s)A(t).$$

Thus the mapping $\rho_R : s_i \to A(s_i)$ is a representation of G, known as the *regular representation*.

By choosing a basis for the vector space we can reformulate the preceding definitions in terms of matrices. A representation of a group G is then a product-preserving map $s \to A(s)$ of G into the group of all $n \times n$ non-singular matrices of complex numbers. The positive integer n is the degree of the representation. However, we must regard two matrix representations $s \to A(s)$ and $s \to B(s)$ as *equivalent* if one is obtained from the other simply by changing the basis of the vector space, i.e. if there exists a non-singular matrix T such that

$$T^{-1}A(s)T = B(s) \quad \text{for every } s \in G.$$

It is easily verified that if $s \to A(s)$ is a matrix representation of degree n of a group G, then $s \to A(s^{-1})^t$ (the transpose of $A(s^{-1})$) is a representation of the same degree, the *contragredient representation*. Furthermore, $s \to \det A(s)$ is a representation of degree 1.

Again, if $\rho : s \to A(s)$ and $\sigma : s \to B(s)$ are matrix representations of a group G, of degrees m and n respectively, then the Kronecker product mapping

$$s \to A(s) \otimes B(s)$$

is also a representation of G, of degree mn, since

$$(A(s) \otimes B(s))(A(t) \otimes B(t)) = A(st) \otimes B(st).$$

We will call this representation simply the *product* of the representations ρ and σ, and denote it by $\rho \otimes \sigma$.

The basic problem of representation theory is to determine all possible representations of a given group. As we will see, all representations may in fact be built up from certain 'irreducible' ones.

Let ρ be a representation of a group G by linear transformations of a vector space V. If a subspace U of V is *invariant* under G, i.e. if

$$\rho(s)U \subseteq U \quad \text{for every } s \in G,$$

then the restrictions to U of the given linear transformations provide a representation ρ_U of G by linear transformations of the vector space U. If it happens that there exists another subspace W invariant under G such that V is the direct sum of U and W, i.e. $V = U + W$ and $U \cap W = \{0\}$, then the representation ρ is completely determined by the representations ρ_U and ρ_W and will be said simply to be their *sum*.

A representation ρ of a group G by linear transformations of a vector space V is said to be *irreducible* if no nontrivial proper subspace of V is invariant under G, and *reducible* otherwise. Evidently any representation of degree 1 is irreducible.

A matrix representation $s \to A(s)$, of degree n, of a group G is reducible if it is equivalent to a representation in which all matrices have the block form

$$\begin{pmatrix} P(s) & Q(s) \\ 0 & R(s) \end{pmatrix},$$

where $P(s)$ is a square matrix of order m, $0 < m < n$. Then $s \to P(s)$ and $s \to R(s)$ are representations of G of degrees m and $n-m$ respectively. The given representation is the sum of these representations if there exists a non-singular matrix T such that

$$T^{-1}A(s)T = \begin{pmatrix} P(s) & 0 \\ 0 & R(s) \end{pmatrix} \quad \text{for every } s \in G.$$

The following theorem of Maschke (1899) reduces the problem of finding all representations of a *finite* group to that of finding all irreducible representations.

Proposition 9 *Every representation of a finite group is (equivalent to) a sum of irreducible representations.*

Proof We give a constructive proof due to Schur. Let $s \to A(s)$, where

$$A(s) = \begin{pmatrix} P(s) & Q(s) \\ 0 & R(s) \end{pmatrix},$$

be a reducible representation of a group G of finite order g. Since the mapping $s \to A(s)$ preserves products, we have

$$P(st) = P(s)P(t), \quad R(st) = R(s)R(t), \quad Q(st) = P(s)Q(t) + Q(s)R(t). \quad (2)$$

The non-singular matrix

$$T = \begin{pmatrix} I & M \\ 0 & I \end{pmatrix}$$

satisfies

$$\begin{pmatrix} P(t) & Q(t) \\ 0 & R(t) \end{pmatrix} T = T \begin{pmatrix} P(t) & 0 \\ 0 & R(t) \end{pmatrix} \quad (3)$$

if and only if

$$MR(t) = P(t)M + Q(t).$$

Take

$$M = g^{-1} \sum_{s \in G} Q(s)R(s^{-1}).$$

Then, by (2),

$$\begin{aligned} P(t)M &= g^{-1} \sum_{s \in G} \{Q(ts) - Q(t)R(s)\}R(s^{-1}) \\ &= g^{-1} \sum_{s \in G} Q(ts)R(s^{-1}t^{-1})R(t) - Q(t) = MR(t) - Q(t), \end{aligned}$$

and hence (3) holds.

Thus the given reducible representation $s \to A(s)$ is the sum of two representations $s \to P(s)$ and $s \to R(s)$ of lower degree. The result follows by induction on the degree. □

Maschke's original proof of Proposition 9 depended on showing that every representation of a finite group is equivalent to a representation by *unitary* matrices. We briefly sketch the argument. Let $\rho : s \to A(s)$ be a representation of a finite group G by linear transformations of a finite-dimensional vector space V. We may suppose V equipped with a positive definite inner product (u, v). It is easily verified that

$$(u, v)_G = g^{-1} \sum_{t \in G} (A(t)u, A(t)v)$$

is also a positive definite inner product on V and that it is invariant under G, i.e.

$$(A(s)u, A(s)v)_G = (u, v)_G \quad \text{for every } s \in G.$$

If U is a subspace of V which is invariant under G, and if $U^\perp$ is the subspace consisting of all vectors $v \in V$ such that $(u, v)_G = 0$ for every $u \in U$, then $U^\perp$ is also invariant under G and V is the direct sum of U and $U^\perp$. Thus ρ is the sum of its restrictions to U and $U^\perp$.

The basic result for irreducible representations is *Schur's lemma*, which comes in two parts:

Proposition 10 (i) *Let $s \to A_1(s)$ and $s \to A_2(s)$ be irreducible representations of a group G by linear transformations of the vector spaces V_1 and V_2. If there exists a linear transformation $T \neq 0$ of V_1 into V_2 such that*

$$T A_1(s) = A_2(s) T \text{ for every } s \in G,$$

then the spaces V_1 and V_2 have the same dimension and T is invertible, so that the representations are equivalent.

(ii) *Let $s \to A(s)$ be an irreducible representation of a group G by linear transformations of a vector space V. A linear transformation T of V has the property*

$$T A(s) = A(s) T \quad \text{for every } s \in G \tag{4}$$

if and only if $T = \lambda I$ for some $\lambda \in \mathbb{C}$.

Proof (i) The image of V_1 under T is a subspace of V_2 which is invariant under the second representation. Since $T \neq 0$ and the representation is irreducible, it must be the whole space: $T V_1 = V_2$. On the other hand, those vectors in V_1 whose image under T is 0 form a subspace of V_1 which is invariant under the first representation. Since $T \neq 0$ and the representation is irreducible, it must contain only the zero vector. Hence distinct vectors of V_1 have distinct images in V_2 under T. Thus T is a one-to-one mapping of V_1 onto V_2.

(ii) By the fundamental theorem of algebra, there exists a complex number λ such that $\det(\lambda I - T) = 0$. Hence $T - \lambda I$ is not invertible. But if T has the property (4), so does $T - \lambda I$. Therefore $T - \lambda I = 0$, by (i) with $A_1 = A_2$. It is obvious that, conversely, (4) holds if $T = \lambda I$. □

Corollary 11 *Every irreducible representation of an abelian group is of degree* 1.

Proof By Proposition 10 (ii) all elements of the group must be represented by scalar multiples of the identity transformation. But such a representation is irreducible only if its degree is 1. □

5 Characters of Arbitrary Finite Groups

By definition, the *trace* of an $n \times n$ matrix $A = (\alpha_{ij})$ is the sum of its main diagonal elements:

$$\mathrm{tr}A = \sum_{i=1}^{n} \alpha_{ii}.$$

It is easily verified that, for any $n \times n$ matrices A, B and any scalars λ, μ, we have

$$\mathrm{tr}(\lambda A + \mu B) = \lambda \mathrm{tr}\, A + \mu \mathrm{tr}\, B,$$
$$\mathrm{tr}(AB) = \mathrm{tr}(BA), \quad \mathrm{tr}(A \otimes B) = (\mathrm{tr}\, A)(\mathrm{tr}\, B).$$

Let $\rho : s \to A(s)$ be a matrix representation of a group G. By the *character* of the representation ρ we mean the mapping $\chi : G \to \mathbb{C}$ defined by

$$\chi(s) = \mathrm{tr}A(s).$$

Since $\mathrm{tr}(T^{-1}AT) = \mathrm{tr}(ATT^{-1}) = \mathrm{tr}A$, equivalent representations have the same character. The significance of characters stems from the converse, which will be proved below.

Clearly the character χ of a representation ρ is a *class function*, i.e.

$$\chi(st) = \chi(ts) \quad \text{for all } s, t \in G.$$

The degree n of the representation ρ is determined by its character χ, since $A(e) = I_n$ and hence $\chi(e) = n$.

If the representation ρ is the sum of two representations ρ' and ρ'', the corresponding characters χ, χ', χ'' evidently satisfy

$$\chi(s) = \chi'(s) + \chi''(s) \quad \text{for every } s \in G.$$

On the other hand, if the representation ρ is the product of the representations ρ' and ρ'', then

$$\chi(s) = \chi'(s)\chi''(s) \quad \text{for every } s \in G.$$

Thus the set of all characters of a group is closed under addition and multiplication. The character of an irreducible representation will be called simply an *irreducible character*.

Let G be a group and ρ a representation of G of degree n with character χ. If s is an element of G of finite order m, then by restriction ρ defines a representation of the cyclic group generated by s. By Proposition 9 and Corollary 11, this representation is equivalent to a sum of representations of degree 1. Thus if S is the matrix representing s, there exists an invertible matrix T such that

$$T^{-1}ST = \mathrm{diag}[\omega_1, \ldots, \omega_n]$$

is a diagonal matrix. Moreover, since

$$T^{-1}S^kT = \operatorname{diag}[\omega_1^k, \ldots, \omega_n^k],$$

$\omega_1, \ldots, \omega_n$ are all m-th roots of unity. Thus

$$\chi(s) = \omega_1 + \cdots + \omega_n$$

is a sum of n m-th roots of unity. Since the inverse of a root of unity ω is its complex conjugate $\bar{\omega}$, it follows that

$$\chi(s^{-1}) = \omega_1^{-1} + \cdots + \omega_n^{-1} = \overline{\chi(s)}.$$

Now let G be a group of finite order g, and let $\rho : s \to A(s)$ and $\sigma : s \to B(s)$ be irreducible matrix representations of G of degrees n and m respectively. For any $n \times m$ matrix C, form the matrix

$$T = \sum_{s \in G} A(s)CB(s^{-1}).$$

Since ts runs through the elements of G at the same time as s,

$$A(t)T = TB(t) \quad \text{for every } t \in G.$$

Therefore, by Schur's lemma, $T = O$ if ρ is not equivalent to σ and $T = \lambda I$ if $\rho = \sigma$. In particular, take C to be any one of the mn matrices which have a single entry 1 and all other entries 0. Then if $A = (\alpha_{ij})$, $B = (\beta_{kl})$, we get

$$\sum_{s \in G} \alpha_{ij}(s)\beta_{kl}(s^{-1}) = \begin{cases} 0 & \text{if } \rho, \sigma \text{ are inequivalent,} \\ \lambda_{jk}\delta_{il} & \text{if } \rho = \sigma, \end{cases}$$

where $\delta_{il} = 1$ or 0 according as $i = l$ or $i \neq l$ ('Kronecker delta'). Since for $(\alpha_{ij}) = (\beta_{ij})$ the left side is unchanged when i is interchanged with k and j with l, we must have $\lambda_{jk} = \lambda\delta_{jk}$. To determine λ set $i = l$, $j = k$ and sum with respect to k. Since the matrices representing s and s^{-1} are inverse, we get $g1 = n\lambda$. Thus

$$\sum_{s \in G} \alpha_{ij}(s)\alpha_{kl}(s^{-1}) = \begin{cases} g/n & \text{if } j = k \text{ and } i = l, \\ 0 & \text{otherwise.} \end{cases}$$

If μ, ν run through an index set for the inequivalent irreducible representations of G, then the relations which have been obtained can be rewritten in the form

$$\sum_{s \in G} \alpha_{ij}^{(\mu)}(s)\alpha_{kl}^{(\nu)}(s^{-1}) = \begin{cases} g/n_\mu & \text{if } \mu = \nu, j = k, i = l, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The *orthogonality relations* (5) for the irreducible matrix elements have several corollaries:

(i) *The functions* $\alpha_{ij}^{(\mu)} : G \to \mathbb{C}$ *are linearly independent.*

For suppose there exist $\lambda_{ij}^{(\mu)} \in \mathbb{C}$ such that

$$\sum_{i,j,\mu} \lambda_{ij}^{(\mu)} \alpha_{ij}^{(\mu)}(s) = 0 \quad \text{for every } s \in G.$$

Multiplying by $\alpha_{kl}^{(\nu)}(s^{-1})$ and summing over all $s \in G$, we get $(g/n_\nu)\lambda_{lk}^{(\nu)} = 0$. Hence every coefficient $\lambda_{lk}^{(\nu)}$ vanishes.

(ii)

$$\sum_{s \in G} \chi_\mu(s)\chi_\nu(s^{-1}) = g\delta_{\mu\nu}. \tag{6}$$

This follows from (5) by setting $i = j, k = l$ and summing over j, l.

(iii) *The irreducible characters χ_μ are linearly independent.*

In fact (iii) follows from (6) in the same way that (i) follows from (5).

The *orthogonality relations* (6) for the irreducible characters enable us to decompose a given representation ρ into irreducible representations. For if $\rho = \oplus m_\mu \rho_\mu$ is a direct sum decomposition of ρ into irreducible components ρ_μ, where the coefficients m_μ are non-negative integers, and if ρ has character χ, then

$$\chi(s) = \sum_\mu m_\mu \chi_\mu(s).$$

Multiplying by $\chi_\nu(s^{-1})$ and summing over all $s \in G$, we deduce from (6) that

$$g^{-1} \sum_{s \in G} \chi(s)\chi_\nu(s^{-1}) = m_\nu. \tag{7}$$

Thus the multiplicities m_ν are uniquely determined by the character χ of the representation ρ. It follows that *two representations are equivalent if and only if they have the same character.*

In the same way we find

$$g^{-1} \sum_{s \in G} \chi(s)\chi(s^{-1}) = \sum_\mu {m_\mu}^2. \tag{8}$$

Hence *a representation ρ with character χ is irreducible if and only if*

$$g^{-1} \sum_{s \in G} \chi(s)\chi(s^{-1}) = 1.$$

The procedure for decomposing a representation into its irreducible components may be applied, in particular, to the regular representation. Evidently the $g \times g$ matrix representing an element s has all its main diagonal elements 0 if $s \neq e$ and all its main diagonal elements 1 if $s = e$. Thus the character χ_R of the regular representation ρ_R is given by

$$\chi_R(e) = g, \quad \chi_R(s) = 0 \quad \text{if } s \neq e.$$

Since $\chi_\nu(e) = n_\nu$ is the degree of the ν-th irreducible representation, it follows from (7) that $m_\nu = n_\nu$. Thus *every irreducible representation is contained in the direct sum decomposition of the regular representation, and moreover each occurs as often as its degree*.

It follows that

$$\sum_\mu {n_\mu}^2 = g, \quad \sum_\mu n_\mu \chi_\mu(s) = 0 \quad \text{if } s \neq e. \tag{9}$$

Thus the total number of functions $\alpha_{ij}^{(\mu)}$ is $\sum_\mu {n_\mu}^2 = g$. Therefore, since they are linearly independent, *every function $\phi : G \to \mathbb{C}$ is a linear combination of functions $\alpha_{ij}^{(\mu)}$ occurring in irreducible matrix representations*.

We show next that *every class function $\phi : G \to \mathbb{C}$ is a linear combination of irreducible characters χ_μ*. By what we have just proved $\phi = \sum_\mu \phi_\mu$, where

$$\phi_\mu = \sum_{i,j=1}^{n_\mu} \lambda_{ij}^{(\mu)} \alpha_{ij}^{(\mu)}$$

and $\lambda_{ij}^{(\mu)} \in \mathbb{C}$. But $\phi(st) = \phi(ts)$ and

$$\phi_\mu(st) = \sum_{i,j,k} \lambda_{ik}^{(\mu)} \alpha_{ij}^{(\mu)}(s) \alpha_{jk}^{(\mu)}(t), \quad \phi_\mu(ts) = \sum_{i,j,k} \lambda_{kj}^{(\mu)} \alpha_{ki}^{(\mu)}(t) \alpha_{ij}^{(\mu)}(s).$$

Since the functions $\alpha_{ij}^{(\mu)}$ are linearly independent, we must have

$$\sum_k \lambda_{ik}^{(\mu)} \alpha_{jk}^{(\mu)}(t) = \sum_k \lambda_{kj}^{(\mu)} \alpha_{ki}^{(\mu)}(t).$$

If we denote by $T^{(\mu)}$ the transpose of the matrix $(\lambda_{ik}^{(\mu)})$, we can rewrite this in the form

$$A^{(\mu)}(t) T^{(\mu)} = T^{(\mu)} A^{(\mu)}(t).$$

Consequently, by Schur's lemma, $T^{(\mu)} = \lambda_\mu I_{n_\mu}$ and hence $\phi_\mu = \lambda_\mu \chi_\mu$. Thus $\phi = \sum_\mu \lambda_\mu \chi_\mu$.

Two elements u, v of a group G are said to be *conjugate* if $v = s^{-1} u s$ for some $s \in G$. It is easily verified that conjugacy is an equivalence relation. Consequently G is the union of pairwise disjoint subsets, called *conjugacy classes*, such that two elements belong to the same subset if and only if they are conjugate. The inverses of all elements in a conjugacy class again form a conjugacy class, the *inverse class*.

In this terminology a function $\phi : G \to \mathbb{C}$ is a class function if and only if $\phi(u) = \phi(v)$ whenever u and v belong to the same conjugacy class. Thus the number of linearly independent class functions is just the number of conjugacy classes in G. Since the characters χ_μ form a basis for the class functions, it follows that *the number of inequivalent irreducible representations is equal to the number of conjugacy classes in the group*.

If a group of order g has r conjugacy classes then, by (9), $g = n_1^2 + \cdots + n_r^2$. Since it is abelian if and only if every conjugacy class contains exactly one element,

i.e. if and only if $r = g$, it follows that *a finite group is abelian if and only if every irreducible representation has degree* 1.

Let $\mathscr{C}_1, \ldots, \mathscr{C}_r$ be the conjugacy classes of the group G and let h_k be the number of elements in $\mathscr{C}_k$ $(k = 1, \ldots, r)$. Changing notation, we will now denote by χ_{ik} the common value of the character of all elements in the k-th conjugacy class in the i-th irreducible representation. Then, since $\chi(s^{-1}) = \overline{\chi(s)}$, the orthogonality relations (6) can be rewritten in the form

$$g^{-1} \sum_{j=1}^{r} h_j \chi_{ij} \overline{\chi_{kj}} = \begin{cases} 1 & \text{if } i = k, \\ 0 & \text{if } i \neq k. \end{cases} \tag{10}$$

Thus the $r \times r$ matrices $A = (\chi_{ik})$, $B = (g^{-1} h_i \overline{\chi_{ki}})$ satisfy $AB = I$. Therefore also $BA = I$, i.e.

$$\sum_{j=1}^{r} \overline{\chi_{ji}} \chi_{jk} = \begin{cases} g/h_k & \text{if } i = k, \\ 0 & \text{if } i \neq k. \end{cases} \tag{11}$$

It may be noted that h_k divides g since, for any $s_k \in \mathscr{C}_k$, g/h_k is the order of the subgroup formed by all elements of G which commute with s_k. We are going to show finally that *the degree of any irreducible representation divides the order of the group*.

Any representation $\rho : s \to A(s)$ of a finite group G may be extended by linearity to the set of all linear combinations of elements of G:

$$\rho\left(\sum_{s \in G} \alpha_s s\right) = \sum_{s \in G} \alpha_s A(s).$$

In particular, let C_k denote the sum of all elements in the k-th conjugacy class $\mathscr{C}_k$ of G. For any $t, u \in G$,

$$u^{-1} s_k u t = t(t^{-1} u^{-1} s_k u t)$$

and hence

$$\rho(C_k) A(t) = \sum_{s \in \mathscr{C}_k} A(st) = \sum_{s \in \mathscr{C}_k} A(ts) = A(t) \rho(C_k).$$

If $\rho = \rho_i$ is an irreducible representation, it follows from Schur's lemma that $\rho_i(C_k) = \lambda_{ik} I_{n_i}$. Moreover, since

$$\operatorname{tr} \rho_i(C_k) = h_k \chi_{ik},$$

where h_k again denotes the number of elements in $\mathscr{C}_k$, we must have $\lambda_{ik} = h_k \chi_{ik} / n_i$. Now let

$$C = \sum_{k=1}^{r} (g/h_k) C_k C_{k'},$$

where $\mathscr{C}_{k'}$ is the conjugacy class inverse to $\mathscr{C}_k$. (Otherwise expressed, $C = \sum_{s,t \in G} sts^{-1}t^{-1}$). Then $\rho_i(C) = \gamma_i I_{n_i}$, where

$$\gamma_i = \sum_{k=1}^{r} (g/h_k) \lambda_{ik} \overline{\lambda_{ik}} = (g/n_i^2) \sum_{k=1}^{r} h_k \chi_{ik} \overline{\chi_{ik}} = (g/n_i)^2,$$

by (10). If $\rho_R(C)$ is the matrix representing C in the regular representation, it follows that there exists an invertible matrix T such that $T^{-1}\rho_R(C)T$ is a diagonal matrix, consisting of the matrices $(g/n_i)^2 I_{n_i}$, repeated n_i times, for every i. In particular, $(g/n_i)^2$ is a root of the characteristic polynomial $\phi(\lambda) = \det(\lambda I_g - \rho_R(C))$ for every i. But $\rho_R(C)$ is a matrix with integer entries and hence the polynomial $\phi(\lambda) = \lambda^g + a_1\lambda^{g-1} + \cdots + a_g$ has integer coefficients $a_1, \ldots, a_g$. The following lemma, already proved in Proposition II.16 but reproved for convenience of reference here, now implies that $(g/n_i)^2$ is an integer and hence that n_i divides g.

Lemma 12 *If $\phi(\lambda) = \lambda^n + a_1\lambda^{n-1} + \cdots + a_n$ is a monic polynomial with integer coefficients $a_1, \ldots, a_n$ and r a rational number such that $\phi(r) = 0$, then r is an integer.*

Proof We can write $r = b/c$, where b and c are relatively prime integers and $c > 0$. Then

$$b^n + a_1 b^{n-1}c + \cdots + a_n c^n = 0$$

and hence c divides b^n. Since c and b have no common prime factor, this implies $c = 1$. □

If we apply the preceding argument to C_k, rather than to C, we see that there exists an invertible matrix T_k such that $T_k^{-1}\rho_R(C_k)T_k$ is a diagonal matrix, consisting of the matrices $(h_k\chi_{ik}/n_i)I_{n_i}$ repeated n_i times, for every i. Thus $h_k\chi_{ik}/n_i$ is a root of the characteristic polynomial $\phi_k(\lambda) = \det(\lambda I_g - \rho_R(C_k))$. Since this is a monic polynomial with integer coefficients, it follows that *$h_k\chi_{ik}/n_i$ is an algebraic integer.*

6 Induced Representations and Examples

Let H be a subgroup of finite index n of a group G, i.e. G is the disjoint union of n left cosets of H:

$$G = s_1 H \cup \cdots \cup s_n H.$$

Also, let there be given a representation $\sigma : t \to A(t)$ of H by linear transformations of a vector space V. The representation $\tilde{\sigma} : s \to \tilde{A}(s)$ of G *induced* by the given representation σ of H is defined in the following way:

Take the vector space $\tilde{V}$ to be the direct sum of n subspaces V_i, where V_i consists of all formal products $s_i \cdot v$ $(v \in V)$ with the rules of combination

$$s_i \cdot (v + v') = s_i \cdot v + s_i \cdot v', \qquad s_i \cdot (\lambda v) = \lambda(s_i \cdot v).$$

Then we set

$$\tilde{A}(s)s_i \cdot v = s_j \cdot A(t)v,$$

where t and s_j are determined from s and s_i by requiring that $t = s_j^{-1}ss_i \in H$. The degree of the induced representation of G is thus n times the degree of the original representation of H.

With respect to a given basis of V let $A(t)$ now denote the matrix representing $t \in H$ and put $A(s) = O$ if $s \in G \backslash H$. If one adopts corresponding bases for each of the subspaces V_i, then the matrix $\tilde{A}(s)$ representing $s \in G$ in the induced representation is the block matrix

$$\tilde{A}(s) = \begin{pmatrix} A(s_1^{-1}ss_1) & A(s_1^{-1}ss_2) & \cdots & A(s_1^{-1}ss_n) \\ A(s_2^{-1}ss_1) & A(s_2^{-1}ss_2) & \cdots & A(s_2^{-1}ss_n) \\ \cdots & \cdots & \cdots & \cdots \\ A(s_n^{-1}ss_1) & A(s_n^{-1}ss_2) & \cdots & A(s_n^{-1}ss_n) \end{pmatrix}.$$

Evidently each row and each column contains exactly one nonzero block. It should be noted also that a different choice of coset representatives $s_i' = s_i t_i$, where $t_i \in H (i = 1, \ldots, n)$, yields an equivalent representation, since

$$\begin{pmatrix} A(t_1)^{-1} & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & A(t_n)^{-1} \end{pmatrix} \tilde{A}(s) \begin{pmatrix} A(t_1) & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & A(t_n) \end{pmatrix}$$
$$= \begin{pmatrix} A(s_1'^{-1}ss_1') & \cdots & A(s_1'^{-1}ss_n') \\ \cdots & \cdots & \cdots \\ A(s_n'^{-1}ss_1') & \cdots & A(s_n'^{-1}ss_n') \end{pmatrix}.$$

Furthermore, changing the order of the cosets corresponds to performing the same permutation on the rows and columns of $\tilde{A}(s)$, and thus also yields an equivalent representation.

It follows that if ψ is the character of the original representation σ of H, then the character $\tilde{\psi}$ of the induced representation $\tilde{\sigma}$ of G is given by

$$\tilde{\psi}(s) = \sum_{i=1}^{n} \psi(s_i^{-1}ss_i),$$

where we set $\psi(s) = 0$ if $s \notin H$. If H is of finite order h, this can be rewritten in the form

$$\tilde{\psi}(s) = h^{-1} \sum_{u \in G} \psi(u^{-1}su), \tag{12}$$

since $\psi(t^{-1}s_i^{-1}ss_it) = \psi(s_i^{-1}ss_i)$ if $t \in H$.

From any representation of a group G we can also obtain a representation of a subgroup H simply by restricting the given representation to H. We will say that the representation of H is *deduced* from that of G. There is a remarkable reciprocity between induced and deduced representations, discovered by Frobenius (1898):

Proposition 13 *Let $\rho : s \to A(s)$ be an irreducible representation of the finite group G and $\sigma : t \to B(t)$ an irreducible representation of the subgroup H. Then the number of times that σ occurs in the representation of H deduced from the representation ρ of G is equal to the number of times that ρ occurs in the representation of G induced by the representation σ of H.*

Proof Let χ denote the character of the representation ρ of G and ψ the character of the representation σ of H. By (7), the number of times that ρ occurs in the complete reduction of the induced representation $\tilde{\sigma}$ is

$$g^{-1}\sum_{s\in G}\tilde{\psi}(s)\chi(s^{-1}) = (gh)^{-1}\sum_{s,u\in G}\psi(u^{-1}su)\chi(s^{-1}).$$

If we put $u^{-1}s^{-1}u = t$, $u^{-1} = v$, then $s^{-1} = v^{-1}tv$ and (t, v) runs through all elements of $G \times G$ at the same time as (s, u). Therefore

$$\begin{aligned} g^{-1}\sum_{s\in G}\tilde{\psi}(s)\chi(s^{-1}) &= (gh)^{-1}\sum_{t,v\in G}\chi(v^{-1}tv)\psi(t^{-1}) \\ &= h^{-1}\sum_{t\in G}\chi(t)\psi(t^{-1}) = h^{-1}\sum_{t\in H}\chi(t)\psi(t^{-1}), \end{aligned}$$

which is the number of times that σ occurs in the complete reduction of the restriction of ρ to H. □

Corollary 14 *Each irreducible representation of a finite group G is contained in a representation induced by some irreducible representation of a given subgroup H.*

A simple, but still significant, application of these results is to the case where the order of the subgroup H is half that of the whole group G. The subgroup H is then necessarily *normal* (as defined in Chapter I, §7) since, for any $v \in G\backslash H$, the elements of $G\backslash H$ form both a single left coset vH and a single right coset Hv. Hence if $s \to A(s)$ is a representation of H, then so also is $s \to A(v^{-1}sv)$, its *conjugate representation*. Since $v^2 \in H$, the conjugate of the conjugate is equivalent to the original representation. Evidently a representation is irreducible if and only if its conjugate representation is irreducible.

On the other hand G has a nontrivial character λ of degree 1, defined by

$$\lambda(s) = 1 \text{ or } -1 \text{ according as } s \in H \text{ or } s \notin H.$$

If χ is an irreducible character of G, then the character $\chi\lambda$ of the product representation is also irreducible, since

$$1 = g^{-1}\sum_{s\in G}\chi(s)\chi(s^{-1}) = \sum_{s\in G}\chi(s)\lambda(s)\chi(s^{-1})\lambda(s^{-1}).$$

Evidently χ and $\chi\lambda$ have the same degree.

If ψ_i is the character of an irreducible representation of H, we will denote by ψ_i^v the character of its conjugate representation. Thus

$$\psi_i^v(s) = \psi_i(v^{-1}sv).$$

The representation and its conjugate are equivalent if and only if $\psi_i^v(s) = \psi_i(s)$ for every $s \in H$.

Consider now the induced representation $\tilde{\psi}_i$ of G. Since H is a normal subgroup, it follows from (12) that

$$\tilde{\psi}_i(s) = \tilde{\psi}_i^v(s) = 0 \quad \text{if } s \in G \backslash H,$$
$$\tilde{\psi}_i(s) = \tilde{\psi}_i^v(s) = \psi_i(s) + \psi_i^v(s) \quad \text{if } s \in H.$$

Hence $\tilde{\psi}_i = \tilde{\psi}_i^v$ and

$$\begin{aligned}\sum_{s\in G} \tilde{\psi}_i(s)\tilde{\psi}_i(s^{-1}) &= \sum_{s\in H} \{\psi_i(s) + \psi_i^v(s)\}\{\psi_i(s^{-1}) + \psi_i^v(s^{-1})\} \\ &= \sum_{s\in H} \psi_i(s)\psi_i(s^{-1}) + \sum_{s\in H} \psi_i^v(s)\psi_i^v(s^{-1}) \\ &\quad + \sum_{s\in H} \{\psi_i(s)\psi_i^v(s^{-1}) + \psi_i(s^{-1})\psi_i^v(s)\}.\end{aligned}$$

Consequently, by the orthogonality relations for H,

$$\sum_{s\in G} \tilde{\psi}_i(s)\tilde{\psi}_i(s^{-1}) = 2h + 2\sum_{s\in H} \psi_i(s)\psi_i^v(s^{-1}).$$

If ψ_i and ψ_i^v are inequivalent, the second term on the right vanishes and we obtain

$$\sum_{s\in G} \tilde{\psi}_i(s)\tilde{\psi}_i(s^{-1}) = g.$$

Thus the induced representation $\tilde{\psi}_i$ of G is irreducible, its degree being twice that of ψ_i.

On the other hand, if ψ_i and ψ_i^v are equivalent, then

$$\sum_{s\in G} \tilde{\psi}_i(s)\tilde{\psi}_i(s^{-1}) = 2g.$$

If $\tilde{\psi}_i = \sum_j m_j \chi_j$ is the decomposition of $\tilde{\psi}_i$ into irreducible characters χ_j of G, it follows from (8) that $\sum_j m_j^2 = 2$. This implies that $\tilde{\psi}_i$ decomposes into two inequivalent irreducible characters of G, say $\tilde{\psi}_i = \chi_k + \chi_l$. We will show that in fact $\chi_l = \chi_k\lambda$.

If $\chi_k(s) = 0$ for all $s \notin H$, then

$$\sum_{s\in H} \chi_k(s)\chi_k(s^{-1}) = \sum_{s\in G} \chi_k(s)\chi_k(s^{-1}) = g = 2h$$

and hence, by the same argument as that just used, the restriction of χ_k to H decomposes into two inequivalent irreducible characters of H. Since the restriction of $\tilde{\psi}_i$ to H is $2\psi_i$, this is a contradiction. We conclude that $\chi_k(s) \neq 0$ for some $s \notin H$, i.e. $\chi_k\lambda \neq \chi_k$. Since χ_k occurs once in the decomposition of $\tilde{\psi}_i$, and $\tilde{\psi}_i(s) = 0$ if $s \notin H$,

$$\begin{aligned} 1 &= g^{-1} \sum_{s \in G} \tilde{\psi}_i(s) \chi_k(s^{-1}) \\ &= g^{-1} \sum_{s \in H} \tilde{\psi}_i(s) \chi_k(s^{-1}) \\ &= g^{-1} \sum_{s \in H} \tilde{\psi}_i(s) \chi_k(s^{-1}) \lambda(s^{-1}) \\ &= g^{-1} \sum_{s \in G} \tilde{\psi}_i(s) \chi_k(s^{-1}) \lambda(s^{-1}). \end{aligned}$$

Thus $\chi_k \lambda$ also occurs once in the decomposition of $\tilde{\psi}_i$, and since $\chi_k \lambda \neq \chi_k$ we must have $\chi_k \lambda = \chi_l$.

In the relation $\sum_i \psi_i(1)^2 = h$, partition the sum into a sum over pairs of distinct conjugate characters and a sum over self-conjugate characters:

$$\Sigma' \{\psi_i(1)^2 + \psi_i^v(1)^2\} + \Sigma'' \psi_i(1)^2 = h.$$

Then for the corresponding characters of G we have

$$\Sigma' \tilde{\psi}_i(1)^2 + \Sigma'' \{\chi_k(1)^2 + \chi_l(1)^2\} = 2\Sigma' \{\psi_i(1)^2 + \psi_i^v(1)^2\} + 2\Sigma'' \psi_i(1)^2 = 2h = g.$$

Since, by Corollary 14, each irreducible character of G appears in the sum on the left, it follows from (9) that each occurs exactly once. Thus we have proved

Proposition 15 *Let the finite group G have a subgroup H of half its order. Then each pair of distinct conjugate characters of H yields by induction a single irreducible character of G of twice the degree, whereas each self-conjugate character of H yields by induction two distinct irreducible characters of G of the same degree, which coincide on H and differ in sign on $G \backslash H$. The irreducible characters of G thus obtained are all distinct, and every irreducible character of G is obtained in this way.*

We will now use Proposition 15 to determine the irreducible characters of several groups of mathematical and physical interest. Let $\mathscr{S}_n$ denote the *symmetric* group consisting of all permutations of the set $\{1, 2, \ldots, n\}$, $\mathscr{A}_n$ the *alternating* group consisting of all even permutations, and C_n the *cyclic* group consisting of all cyclic permutations. Thus $\mathscr{S}_n$ has order $n!$, $\mathscr{A}_n$ has order $n!/2$ and C_n has order n.

The irreducible characters of the abelian group $\mathscr{A}_3 = C_3$ are all of degree 1 and can be arranged as a table in the following way, where ω is a primitive cube root of unity, say $\omega = e^{2\pi i/3} = (-1 + i\sqrt{3})/2$.

	$\mathscr{A}_3$		
	e	(123)	(132)
ψ_1	1	1	1
ψ_2	1	ω	ω^2
ψ_3	1	ω^2	ω

The group $\mathscr{S}_3$ contains $\mathscr{A}_3$ as a subgroup of index 2. The elements of $\mathscr{S}_3$ form three conjugacy classes: $\mathscr{C}_1$ containing only the identity element e, $\mathscr{C}_2$ containing the three

elements (12),(13),(23) of order 2, and $\mathscr{C}_3$ containing the two elements (123),(132) of order 3. The irreducible character ψ_1 of $\mathscr{A}_3$ is self-conjugate and yields two irreducible characters of $\mathscr{S}_3$ of degree 1, the trivial character χ_1 and the sign character $\chi_2 = \chi_1\lambda$. The irreducible characters ψ_2, ψ_3 of $\mathscr{A}_3$ are conjugate and yield a single irreducible character χ_3 of $\mathscr{S}_3$ of degree 2. Thus we obtain the character table:

	$\mathscr{S}_3$		
	$\mathscr{C}_1$	$\mathscr{C}_2$	$\mathscr{C}_3$
χ_1	1	1	1
χ_2	1	-1	1
χ_3	2	0	-1

The elements of $\mathscr{A}_4$ form four conjugacy classes: $\mathscr{C}_1$ containing only the identity element e, $\mathscr{C}_2$ containing the three elements $t_1 = (12)(34)$, $t_2 = (13)(24)$, $t_3 = (14)(23)$ of order 2, $\mathscr{C}_3$ containing four elements of order 3, namely c, ct_1, ct_2, ct_3, where $c = (123)$, and $\mathscr{C}_4$ containing the remaining four elements of order 3, namely $c^2, c^2t_1, c^2t_2, c^2t_3$. Moreover $N = \mathscr{C}_1 \cup \mathscr{C}_2$ is a normal subgroup of order 4, $H = \{e, c, c^2\}$ is a cyclic subgroup of order 3, and

$$\mathscr{A}_4 = HN, \quad H \cap N = \{e\}.$$

If χ is a character of degree 1 of H, then a character ψ of degree 1 of $\mathscr{A}_4$ is defined by

$$\psi(hn) = \chi(h) \quad \text{for all } h \in H, n \in N.$$

Since H is isomorphic to $\mathscr{A}_3$, we obtain in this way three characters ψ_1, ψ_2, ψ_3 of $\mathscr{A}_4$ of degree 1. Since $\mathscr{A}_4$ has order 12, and $12 = 1 + 1 + 1 + 9$, the remaining irreducible character ψ_4 of $\mathscr{A}_4$ has degree 3. The character table of $\mathscr{A}_4$ can be completed by means of the orthogonality relations (11) and has the following form, where again $\omega = (-1 + i\sqrt{3})/2$.

	$\mathscr{A}_4$			
$\lvert\mathscr{C}\rvert$	1	3	4	4
$\mathscr{C}$	$\mathscr{C}_1$	$\mathscr{C}_2$	$\mathscr{C}_3$	$\mathscr{C}_4$
ψ_1	1	1	1	1
ψ_2	1	1	ω	ω^2
ψ_3	1	1	ω^2	ω
ψ_4	3	-1	0	0

The group $\mathscr{S}_4$ contains $\mathscr{A}_4$ as a subgroup of index 2 and $v = (12) \in \mathscr{S}_4 \backslash \mathscr{A}_4$. The elements of $\mathscr{S}_4$ form five conjugacy classes: $\mathscr{C}_1$ containing only the identity element e, $\mathscr{C}_2$ containing six transpositions (jk) $(1 \le j < k \le 4)$, $\mathscr{C}_3$ containing the three elements of order 2 in $\mathscr{A}_4$, $\mathscr{C}_4$ containing eight elements of order 3, and $\mathscr{C}_5$ containing six elements of order 4.

The self-conjugate character ψ_1 of $\mathscr{A}_4$ yields two characters of $\mathscr{S}_4$ of degree 1, the trivial character χ_1 and the sign character $\chi_2 = \chi_1\lambda$; the pair of conjugate characters ψ_2, ψ_3 of $\mathscr{A}_4$ yields an irreducible character χ_3 of $\mathscr{S}_4$ of degree 2; and the

self-conjugate character ψ_4 of $\mathscr{A}_4$ yields two irreducible characters χ_4, χ_5 of $\mathscr{S}_4$ of degree 3. The rows of the character table corresponding to χ_4, χ_5 must have the form

$$\begin{array}{cccc} 3 & x & z & w & y \\ 3 & -x & z & w & -y \end{array}$$

and from the orthogonality relations (11) we obtain $z = -1$, $w = 0$, $xy = -1$. From the orthogonality relations (10) we further obtain $x + y = 0$. Hence $x^2 = 1$ and the complete character table is

	$\mathscr{S}_4$				
$\lvert\mathscr{C}\rvert$	1	6	3	8	6
$\mathscr{C}$	$\mathscr{C}_1$	$\mathscr{C}_2$	$\mathscr{C}_3$	$\mathscr{C}_4$	$\mathscr{C}_5$
χ_1	1	1	1	1	1
χ_2	1	−1	1	1	−1
χ_3	2	0	2	−1	0
χ_4	3	1	−1	0	−1
χ_5	3	−1	−1	0	1

The physical significance of these groups derives from the fact that $\mathscr{A}_4$ (resp. $\mathscr{S}_4$) is isomorphic to the group of all rotations (resp. orthogonal transformations) of $\mathbb{R}^3$ which map a regular tetrahedron onto itself. Similarly $\mathscr{A}_3$ (resp. $\mathscr{S}_3$) is isomorphic to the group of all plane rotations (resp. plane rotations and reflections) which map an equilateral triangle onto itself.

An important property of induced representations was proved by R. Brauer (1953): each character of a finite group is a linear combination with integer coefficients (not necessarily non-negative) of characters induced from characters of elementary subgroups. Here a group is said to be *elementary* if it is the direct product of a group whose order is a power of a prime and a cyclic group whose order is not divisible by that prime.

It may be deduced without difficulty from Brauer's theorem that, if G is a finite group and m the least common multiple of the orders of its elements, then (as had long been conjectured) any irreducible representation of G is equivalent to a representation in the field $\mathbb{Q}(e^{2\pi i/m})$. Green (1955) has shown that Brauer's theorem is actually best possible: if each character of a finite group G is a linear combination with integer coefficients of characters induced from characters of subgroups belonging to some family $\mathscr{F}$, then each elementary subgroup of G is contained in a conjugate of some subgroup in $\mathscr{F}$.

7 Applications

Character theory has turned out to be an invaluable tool in the study of abstract groups. We illustrate this by two results of Burnside (1904) and Frobenius (1901). It is remarkable, first that these applications were found so soon after the development of character theory and secondly that, one century later, there are still no proofs known which do not use character theory.

Lemma 16 *If $\rho: s \to A(s)$ is a representation of degree n of a finite group G, then the character χ of ρ satisfies*

$$|\chi(s)| \leq n \quad \text{for any } s \in G.$$

Moreover, equality holds for some s if and only if $A(s) = \omega I_n$, where $\omega \in \mathbb{C}$.

Proof If $s \in G$ has order m, there exists an invertible matrix T such that

$$T^{-1}A(s)T = \operatorname{diag}[\omega_1, \ldots, \omega_n],$$

where $\omega_1, \ldots, \omega_n$ are m-th roots of unity. Hence $\chi(s) = \omega_1 + \cdots + \omega_n$ and

$$|\chi(s)| \leq |\omega_1| + \cdots + |\omega_n| = n.$$

Moreover $|\chi(s)| = n$ only if $\omega_1, \ldots, \omega_n$ all lie on the same ray through the origin and hence only if they are all equal, since they lie on the unit circle. But then $A(s) = \omega I_n$. □

The *kernel* of the representation ρ is the set K_ρ of all $s \in G$ for which $\rho(s) = I_n$. Evidently K_ρ is a normal subgroup of G. By Lemma 16, K_ρ may be characterized as the set of all $s \in G$ such that $\chi(s) = n$.

Lemma 17 *Let $\rho: s \to A(s)$ be an irreducible representation of degree n of a finite group G, with character χ, and let $\mathscr{C}$ be a conjugacy class of G containing h elements. If h and n are relatively prime then, for any $s \in \mathscr{C}$, either $\chi(s) = 0$ or $A(s) = \omega I_n$ for some $\omega \in \mathbb{C}$.*

Proof Since h and n are relatively prime, there exist integers a, b such that $ah + bn = 1$. Then

$$\chi(s)/n = ah\chi(s)/n + b\chi(s).$$

Since $h\chi(s)/n$ and $\chi(s)$ are algebraic integers, it follows that $\chi(s)/n$ is an algebraic integer. We may assume that $|\chi(s)| < n$, since otherwise the result follows from Lemma 16.

Suppose s has order m. If $(k, m) = 1$, then the conjugacy class containing s^k also has cardinality h and thus $\chi(s^k)/n$ is an algebraic integer, by what we have already proved. Hence

$$\alpha = \prod_k \chi(s^k)/n,$$

where k runs through all positive integers less than m and relatively prime to m, is also an algebraic integer. But $\chi(s^k) = f(\omega^k)$, where ω is a primitive m-th root of unity and

$$f(x) = x^{r_1} + \cdots + x^{r_n}$$

for some non-negative integers $r_1, \ldots, r_n$ less than m. Thus α is a symmetric function of the primitive roots ω^k. Since the *cyclotomic polynomial*

$$\Phi_n(x) = \prod_k (x - \omega^k)$$

has integer coefficients, it follows that $\alpha \in \mathbb{Q}$. Consequently, by Lemma 12, $\alpha \in \mathbb{Z}$.

But $|\alpha| < 1$, since $|\chi(s)| < n$ and $|\chi(s^k)| \leq n$ for every k. Hence $\alpha = 0$, and thus $\chi(s^k) = 0$ for some k with $(k, m) = 1$. If $g(x)$ is the monic polynomial in $\mathbb{Q}[x]$ of least positive degree such that $g(\omega^k) = 0$, then any polynomial in $\mathbb{Q}[x]$ with ω^k as a root must be divisible by $g(x)$. Since we showed in Chapter II, §5 that the cyclotomic polynomial $\Phi_n(x)$ is irreducible over the field $\mathbb{Q}$, it follows that $g(x) = \Phi_n(x)$ and that $\Phi_n(x)$ divides $f(x)$. Hence also $\chi(s) = f(\omega) = 0$. □

Before stating the next result we recall from Chapter I, §7 that a group is said to be *simple* if it contains more than one element and has no nontrivial proper normal subgroup.

Proposition 18 *If a finite group G has a conjugacy class $\mathscr{C}$ of cardinality p^a, for some prime p and positive integer a, then G is not a simple group.*

Proof If $s \in \mathscr{C}$ then, by (9),

$$\sum_\mu n_\mu \chi_\mu(s) = 0.$$

Assume the notation chosen so that χ_1 is the character of the trivial representation. If $\chi_\mu(s) = 0$ for every $\mu > 1$ for which p does not divide n_μ, then the displayed equation has the form $1 + p\zeta = 0$, where ζ is an algebraic integer. Since $-1/p$ is not an integer, this contradicts Lemma 12. Consequently, by Lemma 17, for some $\nu > 1$ we must have $A^{(\nu)}(s) = \omega I_{n_\nu}$, where $\omega \in \mathbb{C}$. The set K_ν of all elements of G which are represented by the identity transformation in the ν-th irreducible representation is a normal subgroup of G. Moreover $K_\nu \neq \{e\}$, since K_ν contains all elements $u^{-1}s^{-1}us$, and $K_\nu \neq G$, since $\nu > 1$. Thus G is not simple. □

Corollary 19 *If G is a group of order $p^a q^b$, where p, q are distinct primes and a, b non-negative integers such that $a + b > 1$, then G is not simple.*

Proof Let $\mathscr{C}_1, \ldots, \mathscr{C}_r$ be the conjugacy classes of G, with $\mathscr{C}_1 = \{e\}$, and let h_k be the cardinality of $\mathscr{C}_k$ $(k = 1, \ldots, r)$. Then h_k divides the order g of G and

$$g = h_1 + \cdots + h_r.$$

Suppose first that $h_j = 1$ for some $j > 1$. Then $\mathscr{C}_j = \{s_j\}$, where s_j commutes with every element of G. Thus the cyclic group H generated by s_j is a normal subgroup of G. Then G is not simple even if $H = G$, since $a + b > 1$ and any proper subgroup of a cyclic group is normal.

Suppose next that $h_k \neq 1$ for every $k > 1$. If G is simple then, by Proposition 18, q divides h_k for every $k > 1$. Since q divides g, it follows that q divides $h_1 = 1$, which is a contradiction. □

It has been shown by Kazarin (1990) that the normal subgroup generated by the elements of the conjugacy class $\mathscr{C}$ in Proposition 18 is *solvable*. Although no proof of Burnside's Proposition 18 is known which does not use character theory, Goldschmidt (1970) and Matsuyama (1973) have given a rather intricate proof of the important Corollary 19 which is purely group theoretic.

The restriction to *two* distinct primes in the statement of Corollary 19 is essential, since the alternating group $\mathscr{A}_5$ of order $60 = 2^2 \cdot 3 \cdot 5$ is simple. It follows at once from Corollary 19, by induction on the order, that any finite group whose order is divisible by at most two distinct primes is *solvable*. P. Hall (1928/1937) has used Corollary 19 to show that a finite group G of order g is solvable if and only if G has a subgroup H of order h for every factorization $g = p^a h$, where $a > 0$ and p is a prime not dividing h.

The second application of group characters, due to Frobenius, has the following statement:

Proposition 20 *If the finite group G has a nontrivial proper subgroup H such that*

$$x^{-1}Hx \cap H = \{e\} \quad \textit{for every } x \in G \backslash H,$$

then G contains a normal subgroup N such that G is the semidirect product of H and N, i.e.

$$G = NH, \quad H \cap N = \{e\}.$$

Proof Obviously $x^{-1}Hx = y^{-1}Hy$ if $y \in Hx$ and the hypotheses imply that $x^{-1}Hx \cap y^{-1}Hy = \{e\}$ if $y \notin Hx$. If g, h are the orders of G, H respectively, it follows that the number of distinct conjugate subgroups $x^{-1}Hx$ (including H itself) is $n = g/h$. Furthermore the number of elements of G which belong to some conjugate subgroup is $n(h-1) + 1 = g - (n-1)$. Thus the set S of elements of G which do not belong to any conjugate subgroup has cardinality $n - 1$.

Let ψ_μ be the character of an irreducible representation of H and $\tilde{\psi}_\mu$ the character of the induced representation of G. By (12) and the hypotheses,

$$\tilde{\psi}_\mu(e) = n\psi_\mu(e), \quad \tilde{\psi}_\mu(s) = 0 \quad \text{if } s \in S, \quad \tilde{\psi}_\mu(s) = \psi_\mu(s) \quad \text{if } s \in H \backslash e.$$

For any fixed μ, form the class function

$$\chi = \tilde{\psi}_\mu - \psi_\mu(e)\{\tilde{\psi}_1 - \chi_1\},$$

where ψ_1 and χ_1 are the characters of the trivial representations of H and G respectively. Then χ is a *generalized character* of G, i.e. $\chi = \sum_\nu m_\nu \chi_\nu$ is a linear combination of irreducible characters χ_ν with integral, but not necessarily non-negative, coefficients m_ν. Moreover

$$\chi(e) = \psi_\mu(e), \quad \chi(s) = \psi_\mu(e) \quad \text{if } s \in S, \quad \chi(s) = \psi_\mu(s) \quad \text{if } s \in H \backslash e.$$

Hence

$$\sum_{s \in H \backslash e} \chi(s)\chi(s^{-1}) = \sum_{s \in H \backslash e} \psi_\mu(s)\psi_\mu(s^{-1}) = h - \psi_\mu(e)^2.$$

Since S has cardinality $n-1$, it follows that

$$\sum_{s\in G}\chi(s)\chi(s^{-1}) = n\{h-\psi_\mu(e)^2\}+\psi_\mu(e)^2+(n-1)\psi_\mu(e)^2 = g.$$

But the formula (8) holds also for generalized characters. Since $\chi(e) > 0$, we conclude that χ is in fact an irreducible character of G. Thus we have an irreducible representation of degree $\chi(e)$ in which the matrices representing elements of S have trace $\chi(e)$. The elements of S must therefore be represented by the unit matrix, i.e. they belong to the kernel K_μ of the representation.

On the other hand, for any $t \in H\backslash e$ we have

$$\sum_{\mu}\psi_\mu(e)\psi_\mu(t) = 0$$

and hence $\psi_\mu(t) \neq \psi_\mu(e)$ for some μ. Thus the intersection of the kernels K_μ for varying μ contains just the elements of S and e. Since K_μ is a normal subgroup of G, it follows that $N = S\cup\{e\}$ is also a normal subgroup. Furthermore, since $H\cap N = \{e\}$, HN has cardinality $hn = g$ and hence $HN = G$. □

A finite group G which satisfies the hypotheses of Proposition 20 is said to be a *Frobenius group*. The subgroup H is said to be a *Frobenius complement* and the normal subgroup N a *Frobenius kernel*. It is readily shown that a finite permutation group is a Frobenius group if and only if it is transitive and no element except the identity fixes more than one symbol. Another characterization follows from Proposition 20: a finite group G is a Frobenius group if and only if it has a nontrivial proper normal subgroup N such that, if $x \in N$ and $x \neq e$, then $xy \neq yx$ for all $y \in G\backslash N$.

Frobenius groups are of some general significance and much is known about their structure. It is easily seen that h divides $n-1$, so that the subgroups H and N have relatively prime orders. It has been shown by Thompson (1959) that the normal subgroup N is a direct product of groups of prime power order. The structure of H is known even more precisely through the work of Burnside (1901) and others.

Applications of group characters of quite a different kind arise in the study of molecular vibrations. We describe one such application within classical mechanics, due to Wigner (1930). However, there are further applications within quantum mechanics, e.g. to the determination of the possible spectral lines in the Raman scattering of light by a substance whose molecules have a particular symmetry group.

A basic problem of classical mechanics deals with the *small oscillations* of a system of particles about an equilibrium configuration. The equations of motion have the form

$$B\ddot{x} + Cx = 0, \tag{13}$$

where $x \in \mathbb{R}^n$ is a vector of generalized coordinates and B, C are positive definite real symmetric matrices. In fact the kinetic energy is $(1/2)\dot{x}^t B\dot{x}$ and, as a first approximation for x near 0, the potential energy is $(1/2)x^t Cx$.

Since B and C are positive definite, there exists (see Chapter V, §4) a non-singular matrix T such that

$$T^t BT = I, \quad T^t CT = D,$$

where D is a diagonal matrix with positive diagonal elements. By the linear transformation $x = Ty$ the equations of motion are brought to the form

$$\ddot{y} + Dy = 0.$$

These 'decoupled' equations can be solved immediately: if

$$y = (\eta_1, \ldots, \eta_n)^t, \quad D = \operatorname{diag}[\omega_1^2, \ldots, \omega_n^2],$$

with $\omega_k > 0$ $(k = 1, \ldots, n)$, then

$$\eta_k = \alpha_k \cos \omega_k t + \beta_k \sin \omega_k t,$$

where $\alpha_k, \beta_k (k = 1, \ldots, n)$ are arbitrary constants of integration. Hence there exist vectors $a_k, b_k \in \mathbb{R}^n$ such that every solution of (13) is a linear combination of solutions of the form

$$a_k \cos \omega_k t, \quad b_k \sin \omega_k t \quad (k = 1, \ldots, n),$$

the so-called *normal modes* of oscillation. The eigenvalues of the matrix $B^{-1}C$ are the squares of the *normal frequencies* $\omega_1, \ldots, \omega_n$.

An important example is the system of particles formed by a molecule of N atoms. Since the displacement of each atom from its equilibrium position is specified by three coordinates, the internal configuration of the molecule without regard to its position and orientation in space may be specified by $n = 3N - 6$ internal coordinates. The determination of the corresponding normal frequencies $\omega_1, \ldots, \omega_n$ may be a formidable task even for moderate values of N. However, the problem is considerably reduced by taking advantage of the symmetry of the molecule.

A *symmetry operation* is an isometry of $\mathbb{R}^3$ which sends the equilibrium position of any atom into the equilibrium position of an atom of the same type. The set of all symmetry operations is clearly a group under composition, the *symmetry group* of the molecule.

For example, the methane molecule CH_4 has four hydrogen atoms at the vertices of a regular tetrahedron and a carbon atom at the centre, from which it follows that the symmetry group of CH_4 is isomorphic to $\mathscr{S}_4$. Similarly, the ammonia molecule NH_3 has three hydrogen atoms and a nitrogen atom at the four vertices of a regular tetrahedron, and hence the symmetry group of NH_3 is isomorphic to $\mathscr{S}_3$.

We return now to the general case. If G is the symmetry group of the molecule, then to each $s \in G$ there corresponds a linear transformation $A(s)$ of the configuration space $\mathbb{R}^n$. Moreover the map $\rho : s \to A(s)$ is a representation of G. Since the kinetic and potential energies are unchanged by a symmetry operation, we have

$$A(s)^t B A(s) = B, \quad A(s)^t C A(s) = C \quad \text{for every } s \in G.$$

It follows that

$$B^{-1} C A(s) = A(s) B^{-1} C \quad \text{for every } s \in G.$$

Assume the notation chosen so that the distinct ω's are $\omega_1, \ldots, \omega_p$ and ω_k occurs m_k times in the sequence $\omega_1, \ldots, \omega_n$ $(k = 1, \ldots, p)$. Thus $n = m_1 + \cdots + m_p$. If V_k is the set of all $v \in \mathbb{R}^n$ such that

$$B^{-1} C v = \omega_k^2 v,$$

then V_k is an m_k-dimensional subspace of $\mathbb{R}^n (k = 1, \ldots, p)$ and $\mathbb{R}^n$ is the direct sum of $V_1, \ldots, V_p$. Moreover each eigenspace V_k is invariant under $A(s)$ for every $s \in G$. Hence, by Maschke's theorem (which holds also for representations in a real vector space), V_k is a direct sum of real-irreducible invariant subspaces. It follows that there exists a real non-singular matrix T such that, for every $s \in G$,

$$T^{-1}A(s)T = \begin{pmatrix} A_1(s) & 0 & \cdots & 0 \\ 0 & A_2(s) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & A_q(s) \end{pmatrix},$$

where $s \to A_k(s)$ is a real-irreducible representation of G, of degree n_k say ($k = 1, \ldots, q$), and

$$T^{-1}B^{-1}CT = \begin{pmatrix} \lambda_1 I_{n_1} & 0 & \cdots & 0 \\ 0 & \lambda_2 I_{n_2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_q I_{n_q} \end{pmatrix}.$$

If the real-irreducible representations $s \to A_k(s)$ ($k = 1, \ldots, q$) are also complex-irreducible, then their degrees and multiplicities can be found by character theory. Thus by decomposing the representation ρ of G into its irreducible components we can determine the degeneracy of the normal frequencies.

We will not consider here the modifications needed when some real-irreducible component is not also complex-irreducible. Also, it should be noted that it may happen 'accidentally' that $\lambda_j = \lambda_k$ for some $j \neq k$.

As a simple illustration of the preceding discussion we consider the ammonia molecule NH_3. Its internal configuration may be described by the six internal coordinates r_1, r_2, r_3 and $\alpha_{23}, \alpha_{31}, \alpha_{12}$, where r_j is the change from its equilibrium value of the distance from the nitrogen atom to the j-th hydrogen atom, and α_{jk} is the change from its equilibrium value of the angle between the rays joining the nitrogen atom to the j-th and k-th hydrogen atoms.

We will determine the character χ of the corresponding representation ρ of the symmetry group $\mathscr{S}_3$. In the notation of the character table previously given for $\mathscr{S}_3$, there is an element $s \in \mathscr{C}_3$ for which the symmetry operation $A(s)$ cyclically permutes r_1, r_2, r_3 and $\alpha_{23}, \alpha_{31}, \alpha_{12}$. Consequently $\chi(s) = 0$ if $s \in \mathscr{C}_3$. Also, there is an element $t \in \mathscr{C}_2$ for which the symmetry operation $A(t)$ interchanges r_1 with r_2 and α_{23} with α_{31}, but fixes r_3 and α_{12}. Consequently $\chi(t) = 2$ if $t \in \mathscr{C}_2$. Since it is obvious that $\chi(e) = 6$, this determines χ and we adjoin it to the character table of $\mathscr{S}_3$:

$\lvert\mathscr{C}\rvert$	1	3	2
$\mathscr{C}$	$\mathscr{C}_1$	$\mathscr{C}_2$	$\mathscr{C}_3$
χ_1	1	1	1
χ_2	1	−1	1
χ_3	2	0	−1
χ	6	2	0

Decomposing the character χ into its irreducible components by means of (7), we obtain $\chi = 2\chi_1 + 2\chi_3$. Since the irreducible representations of $\mathscr{S}_3$ are all real, this means that the configuration space $\mathbb{R}^6$ is the direct sum of four irreducible invariant subspaces, two of dimension 1 and two of dimension 2. Knowing what to look for, we may verify that the one-dimensional subspaces spanned by $r_1 + r_2 + r_3$ and $\alpha_{23} + \alpha_{31} + \alpha_{12}$ are invariant. Also, the two-dimensional subspace formed by all vectors $\mu_1 r_1 + \mu_2 r_2 + \mu_3 r_3$ with $\mu_1 + \mu_2 + \mu_3 = 0$ is invariant and irreducible, and so is the two-dimensional subspace formed by all vectors $\nu_1\alpha_{23} + \nu_2\alpha_{31} + \nu_3\alpha_{12}$ with $\nu_1 + \nu_2 + \nu_3 = 0$. Hence we can find a real non-singular matrix T such that

$$T^{-1}B^{-1}CT = \begin{pmatrix} \lambda_1 I_1 & 0 & 0 & 0 \\ 0 & \lambda_2 I_1 & 0 & 0 \\ 0 & 0 & \lambda_3 I_2 & 0 \\ 0 & 0 & 0 & \lambda_4 I_2 \end{pmatrix}.$$

This shows that the ammonia molecule NH_3 has two nondegenerate normal frequencies and two doubly degenerate normal frequencies.

8 Generalizations

During the past century the character theory of finite groups has been extensively generalized to infinite groups with a topological structure. It may be helpful to give an overview here, without proofs, of this vast development. The reader wishing to pursue some particular topic may consult the references at the end of the chapter.

A *topological group* is a group G with a topology such that the map $(s, t) \to st^{-1}$ of $G \times G$ into G is continuous. Throughout the following discussion we will assume that G is a topological group which, as a topological space, is *locally compact and Hausdorff*, i.e. any two distinct points are contained in open sets whose closures are disjoint compact sets. (A closed set E in a topological space is *compact* if each open cover of E has a finite subcover. In a metric space this is consistent with the definition of sequential compactness in Chapter I, §4.)

Let $\mathscr{C}_0(G)$ denote the set of all continuous functions $f: G \to \mathbb{C}$ such that $f(s) = 0$ for all s outside some compact subset of G (which may depend on f). A map $M: \mathscr{C}_0(G) \to \mathbb{C}$ is said to be a *nonnegative linear functional* if

(i) $M(f_1 + f_2) = M(f_1) + M(f_2)$ for all $f_1, f_2 \in \mathscr{C}_0(G)$,
(ii) $M(\lambda f) = \lambda M(f)$ for all $\lambda \in \mathbb{C}$ and $f \in \mathscr{C}_0(G)$,
(iii) $M(f) \geq 0$ if $f(s) \geq 0$ for every $s \in G$.

It is said to be a *left* (resp. *right*) *Haar integral* if, in addition, it is nontrivial, i.e. $M(f) \neq 0$ for some $f \in \mathscr{C}_0(G)$, and left (resp. right) invariant, i.e.

(iv) $M({}_t f) = M(f)$ for every $t \in G$ and $f \in \mathscr{C}_0(G)$, where ${}_t f(s) = f(t^{-1}s)$, (resp. $M(f_t) = M(f)$ for every $t \in G$ and $f \in \mathscr{C}_0(G)$, where $f_t(s) = f(st)$).

It was shown by Haar (1933) that a left Haar integral exists on any locally compact group; it was later shown to be uniquely determined apart from a positive multiplicative constant. By defining $M^*(f) = M(f^*)$, where $f^*(s) = f(s^{-1})$ for every $s \in G$, it follows that a right Haar integral also exists and is uniquely determined apart from a positive multiplicative constant.

The notions of left and right Haar integral obviously coincide if the group G is abelian, and it may be shown that they also coincide if G is compact or is a semi-simple Lie group.

We now restrict attention to the case of a left Haar integral. It is easily seen that

$$M(\bar{f}) = \overline{M(f)},$$

where $\bar{f}(s) = \overline{f(s)}$ for every $s \in G$. If we set $(f, g) = M(f\bar{g})$, then the usual inner product properties hold:

$$\begin{aligned}(f_1 + f_2, g) &= (f_1, g) + (f_2, g),\\ (\lambda f, g) &= \lambda(f, g),\\ (f, g) &= \overline{(g, f)},\\ (f, f) &\geq 0, \text{ with equality only if } f \equiv 0.\end{aligned}$$

By the *Riesz representation theorem*, there is a unique *positive measure* μ on the σ-algebra $\mathscr{M}$ generated by the compact subsets of G (cf. Chapter XI, §3) such that $\mu(K)$ is finite for every compact set $K \subseteq G$, $\mu(E)$ is the supremum of $\mu(K)$ over all compact $K \subseteq E$ for each $E \in \mathscr{M}$, and

$$M(f) = \int_G f \, \mathrm{d}\mu \quad \text{for every } f \in \mathscr{C}_0(G).$$

The measure μ is necessarily left invariant:

$$\mu(E) = \mu(sE) \quad \text{for all } E \in \mathscr{M} \text{ and } s \in G,$$

where $sE = \{sx : x \in E\}$.

For $p = 1$ or 2, let $L^p(G)$ denote the set of all μ-measurable functions $f : G \to \mathbb{C}$ such that

$$\int_G |f|^p \, \mathrm{d}\mu < \infty.$$

The definition of M can be extended to $L^1(G)$ by setting

$$M(f) = \int_G f \, \mathrm{d}\mu,$$

and the inner product can be extended to $L^2(G)$ by setting

$$(f, g) = \int_G f\bar{g} \, \mathrm{d}\mu.$$

Moreover, with this inner product $L^2(G)$ is a *Hilbert space*. If we define the *convolution product* $f * g$ of $f, g \in L^1(G)$ by

$$f * g(s) = \int_G f(st)g(t^{-1}) \, \mathrm{d}\mu(t),$$

then $L^1(G)$ is a *Banach algebra* and

$$M(f * g) = M(f)M(g) \quad \text{for all } f, g \in L^1(G).$$

A *unitary representation* of G in a Hilbert space $\mathscr{H}$ is a map ρ of G into the set of all linear transformations of $\mathscr{H}$ which maps the identity element e of G into the identity transformation of $\mathscr{H}$:

$$\rho(e) = I,$$

which preserves not only products in G:

$$\rho(st) = \rho(s)\rho(t) \quad \text{for all } s, t \in G,$$

but also inner products in $\mathscr{H}$:

$$(\rho(s)u, \rho(s)v) = (u, v) \quad \text{for all } s \in G \text{ and all } u, v \in \mathscr{H},$$

and for which the map $(s, v) \to \rho(s)v$ of $G \times \mathscr{H}$ into $\mathscr{H}$ is continuous (or, equivalently, for which the map $s \to (\rho(s)v, v)$ of G into $\mathbb{C}$ is continuous at e for every $v \in \mathscr{H}$).

For example, any locally compact group G has a unitary representation ρ in $L^2(G)$, its *regular representation*, defined by

$$(\rho(t)f)(s) = f(t^{-1}s) \quad \text{for all } f \in L^2(G) \text{ and all } s, t \in G.$$

If ρ is a unitary representation of G in a Hilbert space $\mathscr{H}$, and if a closed subspace V of $\mathscr{H}$ is invariant under $\rho(s)$ for every $s \in G$, then so also is its orthogonal complement $V^{\perp}$. The representation ρ is said to be *irreducible* if the only closed subspaces of $\mathscr{H}$ which are invariant under $\rho(s)$ for every $s \in G$ are $\mathscr{H}$ and $\{0\}$. It has been shown by Gelfand and Raikov (1943) that, for any locally compact group G and any $s \in G \backslash e$, there is an irreducible unitary representation ρ of G with $\rho(s) \neq I$.

Consider now the case in which the locally compact group G is abelian. Then any irreducible unitary representation of G is one-dimensional. Hence if we define a *character* of G to be a continuous function $\chi : G \to \mathbb{C}$ such that

(i) $\chi(st) = \chi(s)\chi(t)$ for all $s, t \in G$,
(ii) $|\chi(s)| = 1$ for every $s \in G$,

then every irreducible unitary representation is a character, and vice versa.

If multiplication and inversion of characters are defined pointwise, then the set $\hat{G}$ of all characters of G is again an abelian group, the *dual group* of G. Moreover, we can put a topology on $\hat{G}$ by defining a subset of $\hat{G}$ to be open if it is a union of sets of the form

$$N(\psi, \varepsilon, K) = \{\chi \in \hat{G} : |\chi(s)/\psi(s) - 1| < \varepsilon \text{ for all } s \in K\},$$

where $\psi \in \hat{G}$, $\varepsilon > 0$ and K is a compact subset of G. Then $\hat{G}$ is not only abelian, but also a locally compact topological group.

For each fixed $s \in G$, the map $\hat{s}: \chi \to \chi(s)$ is a character of $\hat{G}$. Moreover the map $s \to \hat{s}$ is one-to-one, by the theorem of Gelfand and Raikov, and every character of $\hat{G}$ is obtained in this way. In fact the *duality theorem* of Pontryagin and van Kampen (1934/5) states that G is isomorphic and homeomorphic to the dual group of $\hat{G}$.

The *Fourier transform* of a function $f \in L^1(G)$ is the function $\hat{f}: \hat{G} \to \mathbb{C}$ defined by

$$\hat{f}(\chi) = \int_G f(s)\overline{\chi(s)}\,\mathrm{d}\mu(s),$$

where μ is the Haar measure on G. If $f_1, f_2 \in L^1(G) \cap L^2(G)$, then $\hat{f}_1, \hat{f}_2 \in L^2(\hat{G})$ and, with a suitable fixed normalization of the Haar measure $\hat{\mu}$ on $\hat{G}$,

$$(f_1, f_2)_G = (\hat{f}_1, \hat{f}_2)_{\hat{G}}.$$

Furthermore, the map $f \to \hat{f}$ can be uniquely extended to a unitary map of $L^2(G)$ onto $L^2(\hat{G})$. This generalizes *Plancherel's theorem* for Fourier integrals on the real line.

If $f = g * h$, where $g, h \in L^1(G)$, then $f \in L^1(G)$ and

$$\hat{f}(\chi) = \hat{g}(\chi)\hat{h}(\chi) \quad \text{for every } \chi \in \hat{G}.$$

If, in addition, $g, h \in L^2(G)$, then $\hat{f} \in L^1(\hat{G})$ and, with the same choice as before for the Haar measure $\hat{\mu}$ on $\hat{G}$, the *Fourier inversion formula* holds:

$$f(s) = \int_{\hat{G}} \hat{f}(\chi)\chi(s)\,\mathrm{d}\hat{\mu}(\chi).$$

The *Poisson summation formula* can also be extended to this general setting. Let H be a closed subgroup of G and let K denote the factor group G/H. If the Haar measures $\mu, \hat{v}$ on $H, \hat{K}$ are suitably chosen then, with appropriate hypotheses on $f \in L^1(G)$,

$$\int_H f(t)\,\mathrm{d}\mu(t) = \int_{\hat{K}} \hat{f}(\psi)\,\mathrm{d}\hat{v}(\psi).$$

We now give some examples (without spelling out the topologies). If $G = \mathbb{R}$ is the additive group of all real numbers, then its characters are the functions $\chi_t: \mathbb{R} \to \mathbb{C}$, with $t \in \mathbb{R}$, defined by

$$\chi_t(s) = e^{its}.$$

In this case G is isomorphic and homeomorphic to $\hat{G}$ itself under the map $t \to \chi_t$. The Haar integral of $f \in L^1(G)$ is the ordinary Lebesgue integral

$$M(f) = \int_{-\infty}^{\infty} f(s)\,ds,$$

the Fourier transform of f is

$$\hat{f}(t) = \int_{-\infty}^{\infty} f(s)e^{-its}ds,$$

and the Fourier inversion formula has the form

$$f(s) = (1/2\pi)\int_{-\infty}^{\infty} \hat{f}(t)e^{its}dt.$$

If $G = \mathbb{Z}$ is the additive group of all integers, then its characters are the functions $\chi_z : \mathbb{Z} \to \mathbb{C}$, with $z \in \mathbb{C}$ and $|z| = 1$, defined by

$$\chi_z(n) = z^n.$$

Thus $\hat{G}$ is the multiplicative group of all complex numbers of absolute value 1. The Haar integral of $f \in L^1(G)$ is

$$M(f) = \sum_{n=-\infty}^{\infty} f(n),$$

the Fourier transform of f is

$$\hat{f}(e^{i\phi}) = \sum_{n=-\infty}^{\infty} f(n)e^{-in\phi},$$

and the Fourier inversion formula has the form

$$f(n) = (1/2\pi)\int_0^{2\pi} \hat{f}(e^{i\phi})e^{in\phi}d\phi.$$

Thus the classical theories of Fourier integrals and Fourier series are just special cases. As another example, let $G = \mathbb{Q}_p$ be the additive group of all p-adic numbers. The characters in this case are the functions $\chi_t : \mathbb{Q}_p \to \mathbb{C}$, with $t \in \mathbb{Q}_p$, defined by

$$\chi_t(s) = e^{2\pi i\lambda(st)},$$

where $\lambda(x) = \sum_{j<0} x_j p^j$ if $x \in \mathbb{Q}_p$ is given by $x = \sum_{j=-\infty}^{\infty} x_j p^j$, $x_j \in \{0, 1, \ldots, p-1\}$ and $x_j = 0$ for all large $j < 0$. Also in this case G is isomorphic and homeomorphic to $\hat{G}$ itself under the map $t \to \chi_t$. If we choose the Haar measure on G so that the measure of the compact set $\mathbb{Z}_p$ of all p-adic integers is 1, then the same choice for $\hat{G}$ is the appropriate one for Plancherel's theorem and the Fourier inversion formula.

Consider next the case in which the group G is compact, but not necessarily abelian. In this case $\mathscr{C}_0(G)$ coincides with the set $\mathscr{C}(G)$ of all continuous functions $f : G \to \mathbb{C}$. The Haar integral is both left and right invariant, and we suppose it normalized so that the integral of the constant 1 has the value 1. Then the integral $M(f)$ of any $f \in \mathscr{C}(G)$, or $L^1(G)$, may be called the *invariant mean* of f.

It may be shown that if ρ is a unitary representation of a compact group G in a Hilbert space $\mathscr{H}$, then $\mathscr{H}$ may be represented as a direct sum $\mathscr{H} = \oplus_\alpha \mathscr{H}_\alpha$ of mutually orthogonal finite-dimensional invariant subspaces $\mathscr{H}_\alpha$ such that, for every α, the restriction of ρ to $\mathscr{H}_\alpha$ is irreducible.

In particular, any irreducible unitary representation of a compact group is finite-dimensional. Consequently it is possible to talk about matrix elements and traces, i.e. characters, of irreducible unitary representations. The orthogonality relations for matrix elements and for characters of irreducible representations of finite groups remain valid for irreducible unitary representations of compact groups if one replaces $g^{-1}\sum_{s\in G} f(s)$ by the invariant mean $M(f)$.

Furthermore, any function $f \in \mathscr{C}(G)$ can be uniformly approximated by finite linear combinations of matrix elements of irreducible unitary representations, and any class function $f \in \mathscr{C}(G)$ can be uniformly approximated by finite linear combinations of characters of irreducible unitary representations. Finally, in the direct sum decomposition of the regular representation into finite-dimensional irreducible unitary representations, each irreducible representation occurs as often as its dimension.

Thus the representation theory of compact groups is completely analogous to that of finite groups. Indeed we may regard the representation theory of finite groups as a special case, since any finite group is compact with the discrete topology and any representation is equivalent to a unitary representation.

An example of a compact group which is neither finite nor abelian is the group $G = SU(2)$ of all 2×2 unitary matrices with determinant 1. The elements of G have the form

$$g = \begin{bmatrix} \gamma & \delta \\ -\bar{\delta} & \bar{\gamma} \end{bmatrix},$$

where γ, δ are complex numbers such that $|\gamma|^2 + |\delta|^2 = 1$. Writing $\gamma = \xi_0 + i\xi_3$, $\delta = \xi_1 + i\xi_2$, we see that topologically $SU(2)$ is homeomorphic to the sphere

$$S^3 = \{x = (\xi_0, \xi_1, \xi_2, \xi_3\} \in \mathbb{R}^4 \colon \xi_0^2 + \xi_1^2 + \xi_2^2 + \xi_3^2 = 1\}$$

and hence is compact and *simply-connected* (i.e. it is path-connected and any closed path can be continuously deformed to a point).

For any integer $n \geq 0$, let V_n denote the vector space of all polynomials $f(z_1, z_2)$ with complex coefficients which are homogeneous of degree n. Writing $z = (z_1, z_2)$, we have

$$zg = (\gamma z_1 - \bar{\delta} z_2, \delta z_1 + \bar{\gamma} z_2).$$

Hence if we define a linear transformation T_g of V_n by $(T_g f)(z) = f(zg)$, then $\rho_n \colon g \to T_g$ is a representation of $SU(2)$ in V_n. It may be shown that this representation is irreducible and is unitary with respect to the inner product

$$\left(\sum_{k=0}^{n} \alpha_k z_1^k z_2^{n-k}, \sum_{k=0}^{n} \beta_k z_1^k z_2^{n-k}\right) = \sum_{k=0}^{n} k!(n-k)!\alpha_k \bar{\beta}_k.$$

Moreover, every irreducible representation of $SU(2)$ is equivalent to ρ_n for some $n \geq 0$.

To determine the character χ_n of ρ_n we observe that any $g \in G$ is conjugate in G to a diagonal matrix

$$t = \begin{pmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{pmatrix},$$

where $\theta \in \mathbb{R}$. If $f_k(z_1, z_2) = z_1^k z_2^{n-k} (0 \leq k \leq n)$, then

$$(T_t f_k)(z_1, z_2) = (e^{i\theta} z_1)^k (e^{-i\theta} z_2)^{n-k} = e^{i(2k-n)\theta} f_k(z_1, z_2).$$

Since the polynomials $f_0, \ldots, f_n$ are a basis for V_n it follows that

$$\chi_n(g) = \chi_n(t) = \sum_{k=0}^{n} e^{i(2k-n)\theta}.$$

Thus $\chi_n(I) = n + 1$, $\chi_n(-I) = (-1)^n (n + 1)$ and

$$\chi_n(g) = \{e^{i(n+1)\theta} - e^{-i(n+1)\theta}\}/\{e^{i\theta} - e^{-i\theta}\} = \sin(n + 1)\theta / \sin\theta \text{ if } g \neq I, -I.$$

From this formula we can easily deduce the decomposition of the product representation $\rho_m \otimes \rho_n$ into irreducible components. Since

$$\begin{aligned} \chi_m(g)\chi_n(g) &= (e^{in\theta} + e^{i(n-2)\theta} + \cdots + e^{-in\theta})\{e^{i(m+1)\theta} - e^{-i(m+1)\theta}\}/\{e^{i\theta} - e^{-i\theta}\} \\ &= \chi_{m+n}(g) + \chi_{m+n-2}(g) + \cdots + \chi_{|m-n|}(g), \end{aligned}$$

we have the *Clebsch–Gordan formula*

$$\rho_m \otimes \rho_n = \rho_{m+n} + \rho_{m+n-2} + \cdots + \rho_{|m-n|}.$$

This formula is the group-theoretical basis for the rule in atomic physics which determines the possible values of the angular momentum when two systems with given angular momenta are coupled.

The complex numbers γ, δ with $|\gamma|^2 + |\delta|^2 = 1$ which specify the matrix $g \in SU(2)$ can be uniquely expressed in the form

$$\gamma = e^{i(\psi+\varphi)/2} \cos\theta/2, \quad \delta = e^{i(\psi-\varphi)/2} \sin\theta/2,$$

where $0 \leq \theta \leq \pi$, $0 \leq \varphi < 2\pi$, $-2\pi \leq \psi < 2\pi$. Then the invariant mean of any continuous function $f : SU(2) \to \mathbb{C}$ is given by

$$M(f) = (1/16\pi^2) \int_{-2\pi}^{2\pi} \int_{0}^{2\pi} \int_{0}^{\pi} f(\theta, \varphi, \psi) \sin\theta \, d\theta \, d\varphi \, d\psi.$$

Another example of a compact group which is neither finite nor abelian is the group $SO(3)$ of all 3×3 real orthogonal matrices with determinant 1. The representations of $SO(3)$ may actually be obtained from those of $SU(2)$, since the two groups are

intimately related. This was already shown in §6 of Chapter I, but another version of the proof will now be given.

The set V of all 2×2 matrices v which are skew-Hermitian and have zero trace,

$$v = \begin{pmatrix} \alpha & \beta \\ -\bar{\beta} & \bar{\alpha} \end{pmatrix}, \quad \text{where } \mathscr{R}\alpha = 0,$$

is a three-dimensional real vector space which may be identified with $\mathbb{R}^3$ by writing $\alpha = i\xi_3$, $\beta = \xi_1 + i\xi_2$. Any $g \in G = SU(2)$ defines a linear transformation $T_g : v \to gvg^{-1}$ of $\mathbb{R}^3$. Moreover T_g is an orthogonal transformation, since if

$$T_g v = v_1 = \begin{pmatrix} \alpha_1 & \beta_1 \\ -\bar{\beta}_1 & \bar{\alpha}_1 \end{pmatrix}$$

then, by the product rule for determinants,

$$|\alpha_1|^2 + |\beta_1|^2 = |\alpha|^2 + |\beta|^2.$$

Hence $\det T_g = \pm 1$. In fact, since T_g is a continuous function of g and $SU(2)$ is connected, we must have $\det T_g = \det T_e = 1$ for every $g \in G$. Thus $T_g \in SO(3)$. Since $T_{gh} = T_g T_h$, the map $g \to T_g$ is a representation of G.

Every element of $SO(3)$ is represented in this way, since

$$\text{if } g_\varphi = \begin{pmatrix} e^{-i\varphi/2} & 0 \\ 0 & e^{i\varphi/2} \end{pmatrix} \quad \text{then } T_{g_\varphi} = B_\varphi = \begin{pmatrix} \cos\varphi & \sin\varphi & 0 \\ -\sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\text{if } h_\theta = \begin{pmatrix} \cos\theta/2 & -\sin\theta/2 \\ \sin\theta/2 & \cos\theta/2 \end{pmatrix} \quad \text{then } T_{h_\theta} = C_\theta = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix},$$

and every $A \in SO(3)$ can be expressed as a product $A = B_\psi C_\theta B_\varphi$, where φ, θ, ψ are *Euler's angles*.

If $T_g = I_3$ is the identity matrix, i.e. if $gv = vg$ for every $v \in V$, then $g = \pm I_2$, since any 2×2 matrix which commutes with both the matrices

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}$$

must be a scalar multiple of the identity matrix. It follows that $SO(3)$ *is isomorphic to the factor group* $SU(2)/\{\pm I_2\}$.

These examples, and higher-dimensional generalizations, can be treated systematically by the theory of Lie groups. A *Lie group* is a group G with the structure of a finite-dimensional real analytic manifold such that the map $(x, y) \to xy^{-1}$ of $G \times G$ into G is real analytic.

Some examples of Lie groups are

(i) a *Euclidean space* $\mathbb{R}^n$ under vector addition;
(ii) an *n-dimensional torus* (or *n-torus*) $\mathbb{T}^n$, i.e. the direct product of n copies of the multiplicative group $\mathbb{T}^1$ of all complex numbers of absolute value 1;

(iii) the *general linear group* $GL(n)$ of all real nonsingular $n \times n$ matrices under matrix multiplication;
(iv) the *orthogonal group* $O(n)$ of all matrices $X \in GL(n)$ such that $X^t X = I_n$;
(v) the *unitary group* $U(n)$ of all complex $n \times n$ matrices X such that $X^* X = I_n$, where X^* is the conjugate transpose of X; ($U(n)$ may be viewed as a subgroup of $GL(2n)$)
(vi) the *unitary symplectic group* $Sp(n)$ of all quaternion $n \times n$ matrices X such that $X^* X = I_n$, where X^* is the conjugate transpose of X. ($Sp(n)$ may be viewed as a subgroup of $GL(4n)$)

The definition implies that any Lie group is a locally compact topological group. The fifth Paris problem of Hilbert (1900) asks for a characterization of Lie groups among all topological groups. A complete solution was finally given by Gleason, Montgomery and Zippin (1953): a topological group can be given the structure of a Lie group if and only if it is *locally Euclidean*, i.e. there is a neighbourhood of the identity which is homeomorphic to $\mathbb{R}^n$ for some n.

The advantage of Lie groups over arbitrary topological groups is that, by replacing them by their Lie algebras, they can be studied by the methods of *linear* analysis.

A real (resp. complex) *Lie algebra* is a finite-dimensional real (resp. complex) vector space L with a map $(u, v) \to [u, v]$ of $L \times L$ into L, which is linear in u and in v and has the properties

(i) $[v, v] = 0$ for every $v \in L$,
(ii) $[u, [v, w]] + [v, [w, u]] + [w, [u, v]] = 0$ for all $u, v, w \in L$. (Jacobi identity)

It follows from (i) and the linearity of the bracket product that

$$[u, v] + [v, u] = 0 \quad \text{for all } u, v \in L.$$

An example of a real (resp. complex) Lie algebra is the vector space $\mathfrak{gl}(n, \mathbb{R})$ (resp. $\mathfrak{gl}(n, \mathbb{C})$) of all $n \times n$ real (resp. complex) matrices X with $[X, Y] = XY - YX$. Other examples are easily constructed as subalgebras.

A *Lie subalgebra* of a Lie algebra L is a vector subspace M of L such that $u \in M$ and $v \in M$ imply $[u, v] \in M$. Some Lie subalgebras of $\mathfrak{gl}(n, \mathbb{C})$ are

(i) the set A_n of all $X \in \mathfrak{gl}(n+1, \mathbb{C})$ with $\operatorname{tr} X = 0$,
(ii) the set B_n of all $X \in \mathfrak{gl}(2n+1, \mathbb{C})$ such that $X^t + X = 0$,
(iii) the set C_n of all $X \in \mathfrak{gl}(2n, \mathbb{C})$ such that $X^t J + JX = 0$, where

$$J = \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix},$$

(iv) the set D_n of all $X \in \mathfrak{gl}(2n, \mathbb{C})$ such that $X^t + X = 0$.

The manifold structure of a Lie group G implies that with each $s \in G$ there is associated a real vector space, the *tangent space* at s. The group structure of the Lie group G implies that the tangent space at the identity e of G is a real Lie algebra, which will be denoted by $L(G)$. For example, if $G = GL(n)$ then $L(G) = \mathfrak{gl}(n, \mathbb{R})$. The properties of Lie groups are mirrored by those of their Lie algebras in the following way.

For every real Lie algebra L, there is a simply-connected Lie group $\tilde{G}$ such that $L(\tilde{G}) = L$. Moreover, $\tilde{G}$ is uniquely determined up to isomorphism by L. A connected Lie group G has $L(G) = L$ if and only if G is isomorphic to a factor group $\tilde{G}/D$, where D is a discrete subgroup of the centre of $\tilde{G}$.

A *Lie subgroup* of a Lie group G is a real analytic submanifold H of G which is also a Lie group under the restriction to H of the group structure on G. It may be shown that a subgroup H of a Lie group G is a Lie subgroup if it is a closed subset of G, and is a connected Lie subgroup if and only if it is path-connected. Thus any closed subgroup of $GL(n)$ is a Lie group.

If H is a Lie subgroup of the Lie group G, then $L(H)$ is a Lie subalgebra of $L(G)$. Moreover, if M is a Lie subalgebra of $L(G)$, there is a unique connected Lie subgroup H of G such that $L(H) = M$.

If G_1, G_2 are Lie groups, then a map $f : G_1 \to G_2$ is a *Lie group homomorphism* if it is an analytic map, regarding G_1, G_2 as manifolds, and a homomorphism, regarding G_1, G_2 as groups. It may be shown that any continuous map $f : G_1 \to G_2$ which is a group homomorphism is actually a Lie group homomorphism. (It follows that a locally Euclidean topological group can be given the structure of a Lie group in only one way.)

If L_1, L_2 are Lie algebras, then a map $T : L_1 \to L_2$ is a *Lie algebra homomorphism* if it is linear and $T[u, v] = [Tu, Tv]$ for all $u, v \in L_1$. If G_1, G_2 are Lie groups and if $f : G_1 \to G_2$ is a Lie group homomorphism, then the derivative of f at the identity, $f'(e) : L(G_1) \to L(G_2)$, is a Lie algebra homomorphism. Moreover, if G_1 is connected then distinct Lie group homomorphisms give rise to distinct Lie algebra homomorphisms, and if G_1 is simply-connected then every Lie algebra homomorphism $L(G_1) \to L(G_2)$ arises from some Lie group homomorphism. (In particular, the representations of a connected Lie group are determined by the representations of its Lie algebra.)

A Lie algebra L is *abelian* if $[u, v] = 0$ for all $u, v \in L$. A connected Lie group is abelian if and only if its Lie algebra is abelian. Since the Euclidean space $\mathbb{R}^n$ is a simply-connected Lie group with an n-dimensional abelian Lie algebra, it follows that any n-dimensional connected abelian Lie group is isomorphic to a direct product $\mathbb{R}^{n-k} \times \mathbb{T}^k$ (where $\mathbb{T}^k$ is a k-torus) for some k such that $0 \leq k \leq n$.

An *ideal* of a Lie algebra L is a vector subspace M of L such that $u \in L$ and $v \in M$ imply $[u, v] \in M$. A connected Lie subgroup H of a connected Lie group G is a normal subgroup if and only if $L(H)$ is an ideal of $L(G)$.

A Lie algebra L is *simple* if it has no ideals except $\{0\}$ and L and is not one-dimensional, and *semisimple* if it has no abelian ideal except $\{0\}$. It may be shown that a Lie algebra is semisimple if and only if it is the direct sum of finitely many ideals, each of which is a simple Lie algebra.

A Lie group is *semisimple* if it is connected and has no connected abelian normal Lie subgroup except $\{e\}$. It follows that a connected Lie group G is semisimple if and only if its Lie algebra $L(G)$ is semisimple.

We turn our attention now to compact Lie groups. It may be shown that a compact topological group can be given the structure of a Lie group if and only if it is finite-dimensional and locally connected. Furthermore, a compact Lie group is isomorphic to a closed subgroup of $GL(n)$ for some n. Other basic results are:

(i) a compact Lie group, and even any compact topological group, has only finitely many connected components;

(ii) a connected compact Lie group is abelian if and only if it is an n-torus $\mathbb{T}^n$ for some n;

(iii) a semisimple connected compact Lie group G has a finite centre. Moreover the simply-connected Lie group $\tilde{G}$ such that $L(\tilde{G}) = L(G)$ is not only semisimple but also compact;

(iv) an arbitrary connected compact Lie group G has the form $G = ZH$, where Z, H are connected compact Lie subgroups, H is semisimple and Z is the component of the centre of G which contains the identity e.

These results essentially reduce the classification of arbitrary compact Lie groups to the classification of those which are semisimple and simply-connected. It may be shown that the latter are in one-to-one correspondence with the semisimple *complex* Lie algebras. Since a semisimple Lie algebra is a direct sum of finitely many simple Lie algebras, we are thus reduced to the classification of the simple complex Lie algebras. The miracle is that these can be completely enumerated: the non-isomorphic simple complex Lie algebras consist of the four infinite families $A_n (n \geq 1)$, $B_n (n \geq 2)$, $C_n (n \geq 3)$, $D_n (n \geq 4)$, of dimensions $n(n+2)$, $n(2n+1)$, $n(2n+1)$, $n(2n-1)$ respectively, and five *exceptional* Lie algebras G_2, F_4, E_6, E_7, E_8 of dimensions 14, 52, 78, 133, 248 respectively.

To the simple complex Lie algebra A_n corresponds the compact Lie group $SU(n+1)$ of all matrices in $U(n+1)$ with determinant 1; to B_n corresponds the compact Lie group $SO(2n+1)$ of all matrices in $O(2n+1)$ with determinant 1; to C_n corresponds the compact Lie group $Sp(n)$ (whose matrices all have determinant 1), and to D_n corresponds the compact Lie group $SO(2n)$ of all matrices in $O(2n)$ with determinant 1. The groups $SU(n)$ and $Sp(n)$ are simply-connected if $n \geq 2$, whereas $SO(n)$ is connected but has index 2 in its simply-connected covering group $Spin(n)$ if $n \geq 5$. The compact Lie groups corresponding to the five exceptional simple complex Lie algebras are all related to the algebra of *octonions* or Cayley numbers.

Space does not permit consideration here of the methods by which this classification has been obtained, although the methods are just as significant as the result. Indeed they provide a uniform approach to many problems involving the classical groups, giving explicit formulas for the invariant mean and for the characters of all irreducible representations. There is also a notable connection with *groups generated by reflections*.

The classification of arbitrary semisimple Lie groups reduces similarly to the classification of simple *real* Lie algebras, which have also been completely enumerated. The irreducible unitary representations of non-compact semisimple Lie groups have been extensively studied, notably by Harish-Chandra. However, the non-compact case is essentially more difficult than the compact, since any nontrivial representation is infinite-dimensional, and the results are still incomplete. Much of the motivation for this work has come from elementary particle physics where, in the original formulation of Wigner (1939), a particle (specified by its mass and spin) corresponds to an irreducible unitary representation of the inhomogeneous Lorentz group.

9 Further Remarks

The history of Legendre's conjectures on primes in arithmetic progressions is described in Vol. I of Dickson [13]. Dirichlet's original proof is contained in [33], pp. 313–342. Although no simple general proof of Dirichlet's theorem is known, simple proofs have been given for the existence of infinitely many primes congruent to 1 mod m; see Sedrakian and Steinig [41].

If all arithmetic progressions $a, a+m, \ldots$ with $(a,m) = 1$ contain a prime, then they all contain infinitely many, since for any $k > 1$ the arithmetic progression $a+m^k, a+2m^k, \ldots$ contains a prime.

It may be shown that any finite abelian group G is *isomorphic* to its dual group $\hat{G}$ (although not in a canonical way) by expressing G as a direct product of cyclic groups; see, for example, W. & F. Ellison [15].

In the final step of the proof of Proposition 7 we have followed Bateman [3]. Other proofs that $L(1,\chi) \neq 0$ for every $\chi \neq \chi_1$, which do not use Proposition 6, are given in Hasse [21]. The functional equation for Dirichlet L-functions was first proved by Hurwitz (1882). For proofs of some of the results stated at the end of §3, see Bach and Sorenson [1], Davenport [12], W. & F. Ellison [15] and Prachar [40]. Funakura [18] characterizes Dirichlet L-functions by means of their analytic properties.

The history of the theory of group representations and group characters is described in Curtis [10]. More complete expositions of the subject than ours are given by Serre [42], Feit [16], Huppert [27], and Curtis and Reiner [11]. The proof given here that the degree of an irreducible representation divides the order of the group is not Frobenius' original proof. It first appeared in a footnote of a paper by Schur (1904) on projective representations, where it is attributed to Frobenius. Zassenhaus [50] gives an interpretation in terms of *Casimir operators*.

A character-free proof of Corollary 19 is given in Gagen [19]. P. Hall's theorem is proved in Feit [16], for example. Frobenius groups are studied further in Feit [16] and Huppert [27].

For physical and chemical applications of group representations, see Cornwell [9], Janssen [29], Meijer [36], Birman [4] and Wilson *et al.* [48].

Dym and McKean [14] give an outward-looking introduction to the classical theory of Fourier series and integrals. The formal definition of a topological group is due to Schreier (1926). The Haar integral is discussed by Nachbin [37]. General introductions to abstract harmonic analysis are given by Weil [46], Loomis [34] and Folland [17]. More detailed information on topological groups and their representations is contained in Pontryagin [39], Hewitt and Ross [23] and Gurarii [20]. A simple proof that the additive group $\mathbb{Q}_p$ of all p-adic numbers is isomorphic to its dual group is given by Washington [45]. In the adelic approach to algebraic number theory this isomorphism lies behind the functional equation of the Riemann zeta function; see, for example, Lang [31].

For Hilbert's fifth problem, see Yang [49] and Hirschfeld [24]. The correspondence between Lie groups and Lie algebras was set up by Sophus Lie (1873–1893) in a purely local way, i.e. between neighbourhoods of the identity in the Lie group and of zero in the Lie algebra. Over half a century elapsed before the correspondence was made global by Cartan, Pontryagin and Chevalley. A basic property of solvable Lie algebras

was established by Lie, but we owe to Killing (1888–1890) the remarkable classification of simple complex Lie algebras. Some gaps and inaccuracies in Killing's pioneering work were filled and corrected in the thesis of Cartan (1894). The classification of simple real Lie algebras is due to Cartan (1914). The representation theory of semisimple Lie algebras and compact semisimple Lie groups is the creation of Cartan (1913) and Weyl (1925–7). The introduction of groups generated by reflections is due to Weyl.

For the theory of Lie groups, see Chevalley [7], Warner [44], Varadarajan [43], Helgason [22] and Barut and Raczka [2]. The last reference also has information on representations of noncompact Lie groups and applications to quantum theory. The purely algebraic theory of Lie algebras is discussed by Jacobson [28] and Humphreys [25]. Niederle [38] gives a survey of the applications of the exceptional Lie algebras and Lie superalgebras in particle physics. Groups generated by reflections are treated by Humphreys [26], Bourbaki [5] and Kac [30], while Cohen [8] gives a useful overview.

The character theory of locally compact abelian groups, whose roots lie in Dirichlet's theorem on primes in arithmetic progressions, has given something back to number theory in the adelic approach to algebraic number fields; see the thesis of Tate, reproduced (pp. 305–347) in Cassels and Fröhlich [6], Lang [31] and Weil [47]. For a broad historical perspective and future plans, see Mackey [35] and Langlands [32].

10 Selected References

[1] E. Bach and J. Sorenson, Explicit bounds for primes in residue classes, *Math. Comp.* **65** (1996), 1717–1735.

[2] A.O. Barut and R. Raczka, *Theory of group representations and applications*, 2nd ed., Polish Scientific Publishers, Warsaw, 1986.

[3] P.T. Bateman, A theorem of Ingham implying that Dirichlet's L-functions have no zeros with real part one, *Enseign. Math.* **43** (1997), 281–284.

[4] J.L. Birman, *Theory of crystal space groups and lattice dynamics*, Springer-Verlag, Berlin, 1984.

[5] N. Bourbaki, *Groupes et algèbres de Lie*: *Chapitres 4,5 et 6*, Masson, Paris, 1981.

[6] J.W.S. Cassels and A. Fröhlich (ed.), *Algebraic number theory*, Academic Press, London, 1967.

[7] C. Chevalley, *Theory of Lie groups I*, Princeton University Press, Princeton, 1946. [Reprinted, 1999]

[8] A.M. Cohen, Coxeter groups and three related topics, *Generators and relations in groups and geometries* (ed. A. Barlotti *et al.*), pp. 235–278, Kluwer, Dordrecht, 1991.

[9] J.F. Cornwell, *Group theory in physics*, 3 vols., Academic Press, London, 1984–1989.

[10] C.W. Curtis, *Pioneers of representation theory*: *Frobenius, Burnside, Schur, and Brauer*, American Mathematical Society, Providence, R.I., 1999.

[11] C.W. Curtis and I. Reiner, *Methods of representation theory*, 2 vols., Wiley, New York, 1990.

[12] H. Davenport, *Multiplicative number theory*, 3rd ed. revised by H.L. Montgomery, Springer-Verlag, New York, 2000.

[13] L.E. Dickson, *History of the theory of numbers*, 3 vols., reprinted Chelsea, New York, 1966.

[14] H. Dym and H.P. McKean, *Fourier series and integrals*, Academic Press, Orlando, FL, 1972.

[15] W. Ellison and F. Ellison, *Prime numbers*, Wiley, New York, 1985.
[16] W. Feit, *Characters of finite groups*, Benjamin, New York, 1967.
[17] G.B. Folland, *A course in abstract harmonic analysis*, CRC Press, Boca Raton, FL, 1995.
[18] T. Funakura, On characterization of Dirichlet L-functions, *Acta Arith*. **76** (1996), 305–315.
[19] T.M. Gagen, *Topics in finite groups*, London Mathematical Society Lecture Note Series **16**, Cambridge University Press, 1976.
[20] V.P. Gurarii, *Group methods in commutative harmonic analysis*, English transl. by D. and S. Dynin, Encyclopaedia of Mathematical Sciences **25**, Springer-Verlag, Berlin, 1998.
[21] H. Hasse, *Vorlesungen über Zahlentheorie*, 2nd ed., Springer-Verlag, Berlin, 1964.
[22] S. Helgason, *Differential geometry, Lie groups and symmetric spaces*, Academic Press, New York, 1978.
[23] E. Hewitt and K.A. Ross, *Abstract harmonic analysis*, 2 vols., Springer-Verlag, Berlin, 1963/1970. [Corrected reprint of Vol. I, 1979]
[24] J. Hirschfeld, The nonstandard treatment of Hilbert's fifth problem, *Trans. Amer. Math. Soc.* **321** (1990), 379–400.
[25] J.E. Humphreys, *Introduction to Lie algebras and representation theory*, Springer-Verlag, New York, 1972.
[26] J.E. Humphreys, *Reflection groups and Coxeter groups*, Cambridge University Press, Cambridge, 1990.
[27] B. Huppert, *Character theory of finite groups,* de Gruyter, Berlin, 1998.
[28] N. Jacobson, *Lie algebras*, Interscience, New York, 1962.
[29] T. Janssen, *Crystallographic groups*, North-Holland, Amsterdam, 1973.
[30] V.G. Kac, *Infinite dimensional Lie Algebras*, corrected reprint of 3rd ed., Cambridge University Press, Cambridge, 1995.
[31] S. Lang, *Algebraic number theory*, 2nd ed., Springer-Verlag, New York, 1994.
[32] R.P. Langlands, Representation theory: its rise and its role in number theory, *Proceedings of the Gibbs symposium* (ed. D.G. Caldi and G.D. Mostow), pp. 181–210, Amer. Math. Soc., Providence, Rhode Island, 1990.
[33] G. Lejeune-Dirichlet, *Werke*, reprinted in one volume, Chelsea, New York, 1969.
[34] L.H. Loomis, *An introduction to abstract harmonic analysis*, Van Nostrand, New York, 1953.
[35] G.W. Mackey, Harmonic analysis as the exploitation of symmetry - a historical survey, *Bull. Amer. Math. Soc.* (*N.S.*) **3** (1980), 543–698. [Reprinted, with related articles, in G.W. Mackey, *The scope and history of commutative and noncommutative harmonic analysis*, American Mathematical Society, Providence, R.I., 1992]
[36] P.H. Meijer (ed.), *Group theory and solid state physics*: *a selection of papers*, Vol. 1, Gordon and Breach, New York, 1964.
[37] L. Nachbin, *The Haar integral*, reprinted, Krieger, Huntington, New York, 1976.
[38] J. Niederle, The unusual algebras and their applications in particle physics, *Czechoslovak J. Phys. B* **30** (1980), 1–22.
[39] L.S. Pontryagin, *Topological groups*, English transl. of 2nd ed. by A. Brown, Gordon and Breach, New York, 1966. [Russian original, 1954]
[40] K. Prachar, *Primzahlverteilung*, Springer-Verlag, Berlin, 1957.
[41] N. Sedrakian and J. Steinig, A particular case of Dirichlet's theorem on arithmetic progressions, *Enseign. Math.* **44** (1998), 3–7.
[42] J.-P. Serre, *Linear representations of finite groups*, Springer-Verlag, New York, 1977.
[43] V.S. Varadarajan, *Lie groups, Lie algebras and their representations*, corrected reprint, Springer-Verlag, New York, 1984.
[44] F.W. Warner, *Foundations of differentiable manifolds and Lie groups*, corrected reprint, Springer-Verlag, New York, 1983.
[45] L. Washington, On the self-duality of Q_p, *Amer. Math. Monthly* **81** (1974), 369–371.

[46] A. Weil, *L'integration dans les groupes topologiques et ses applications*, 2nd ed., Hermann, Paris, 1953.
[47] A. Weil, *Basic number theory*, 2nd ed., Springer-Verlag, Berlin, 1973.
[48] E.B. Wilson, J.C. Decius and P.C. Cross, *Molecular vibrations*, McGraw-Hill, New York, 1955.
[49] C.T. Yang, Hilbert's fifth problem and related problems on transformation groups, *Mathematical developments arising from Hilbert problems* (ed. F.E. Browder), pp. 142–146, Amer. Math. Soc., Providence, R.I., 1976.
[50] H. Zassenhaus, An equation for the degrees of the absolutely irreducible representations of a group of finite order, *Canad. J. Math*. **2** (1950), 166–167.

XI

Uniform Distribution and Ergodic Theory

A trajectory of a system which is evolving with time may be said to be 'recurrent' if it keeps returning to any neighbourhood, however small, of its initial point, and 'dense' if it passes arbitrarily near to every point. It may be said to be 'uniformly distributed' if the proportion of time it spends in any region tends asymptotically to the ratio of the volume of that region to the volume of the whole space. In the present chapter these notions will be made precise and some fundamental properties derived. The subject of dynamical systems has its roots in mechanics, but we will be particularly concerned with its applications in number theory.

1 Uniform Distribution

Before introducing our subject, we establish the following interesting result:

Lemma 0 *Let $J = [a, b]$ be a compact interval and $f_n : J \to \mathbb{R}$ a sequence of non-decreasing functions. If $f_n(t) \to f(t)$ for every $t \in J$ as $n \to \infty$, where $f : J \to \mathbb{R}$ is a continuous function, then $f_n(t) \to f(t)$* uniformly *on J.*

Proof Evidently f is also nondecreasing. Furthermore, since J is compact, f is uniformly continuous on J. It follows that, for any $\varepsilon > 0$, there is a subdivision $a = t_0 < t_1 < \cdots < t_m = b$ such that

$$f(t_k) - f(t_{k-1}) < \varepsilon \ (k = 1, \ldots, m).$$

We can choose a positive integer p so that, for all $n > p$,

$$|f_n(t_k) - f(t_k)| < \varepsilon \ (k = 0, 1, \ldots, m).$$

If $t \in J$, then $t \in [t_{k-1}, t_k]$ for some $k \in \{1, \ldots, m\}$. Hence

$$f_n(t) - f(t) \le f_n(t_k) - f(t_k) + f(t_k) - f(t_{k-1}) < 2\varepsilon$$

and similarly

$$f_n(t) - f(t) \ge f_n(t_{k-1}) - f(t_{k-1}) + f(t_{k-1}) - f(t_k) > -2\varepsilon.$$

Thus $|f_n(t) - f(t)| < 2\varepsilon$ for every $t \in J$ if $n > p$. □

W.A. Coppel, *Number Theory: An Introduction to Mathematics*, Universitext,
DOI: 10.1007/978-0-387-89486-7_11, © Springer Science + Business Media, LLC 2009

For any real number ξ, let $\lfloor \xi \rfloor$ denote again the greatest integer $\leq \xi$ and let

$$\{\xi\} = \xi - \lfloor \xi \rfloor$$

denote the *fractional part* of ξ. We are going to prove that, if ξ is irrational, then the sequence $(\{n\xi\})$ of the fractional parts of the multiples of ξ is *dense* in the unit interval $I = [0, 1]$, i.e. every point of I is a limit point of the sequence.

It is sufficient to show that the points $z_n = e^{2\pi i n\xi} (n = 1, 2, \ldots)$ are dense on the unit circle. Since ξ is irrational, the points z_n are all distinct and $z_n \neq \pm 1$. Consequently they have a limit point on the unit circle. Thus, for any given $\varepsilon > 0$, there exist positive integers m, r such that

$$|z_{m+r} - z_m| < \varepsilon.$$

But

$$|z_{m+r} - z_m| = |z_r - 1| = |z_{n+r} - z_n| \quad \text{for every } n \in \mathbb{N}.$$

If we write $z_r = e^{2\pi i\theta}$, where $0 < \theta < 1$, then $z_{kr} = e^{2\pi ik\theta} (k = 1, 2, \ldots)$. Define the positive integer N by $1/(N + 1) < \theta < 1/N$. Then the points $z_r, z_{2r}, \ldots, z_{Nr}$ follow one another in order on the unit circle and every point of the unit circle is distant less than ε from one of these points.

It may be asked if the sequence $(\{n\xi\})$ is not only dense in I, but also spends 'the right amount of time' in each subinterval of I. To make the question precise we introduce the following definition:

A sequence (ξ_n) of real numbers is said to be *uniformly distributed mod* 1 if, for all α, β with $0 \leq \alpha < \beta \leq 1$,

$$\varphi_{\alpha,\beta}(N)/N \to \beta - \alpha \quad \text{as } N \to \infty,$$

where $\varphi_{\alpha,\beta}(N)$ is the number of positive integers $n \leq N$ such that $\alpha \leq \{\xi_n\} < \beta$.

In this definition we need only require that $\varphi_{0,\alpha}(N)/N \to \alpha$ for every $\alpha \in (0, 1)$, since

$$\varphi_{\alpha,\beta}(N) = \varphi_{0,\beta}(N) - \varphi_{0,\alpha}(N)$$

and hence

$$|\varphi_{\alpha,\beta}(N)/N - (\beta - \alpha)| \leq |\varphi_{0,\beta}(N)/N - \beta| + |\varphi_{0,\alpha}(N)/N - \alpha|.$$

It follows from Lemma 0, with $f_n(t) = \varphi_{0,t}(n)/n$ and $f(t) = t$, that the sequence (ξ_n) is uniformly distributed mod 1 if and only if

$$\varphi_{\alpha,\beta}(N)/N \to \beta - \alpha \quad \text{as } N \to \infty$$

uniformly for all α, β with $0 \leq \alpha < \beta \leq 1$.

It was first shown by Bohl (1909) that, if ξ is irrational, the sequence $(n\xi)$ is uniformly distributed mod 1 in the sense of our definition. Later Weyl (1914,1916) established this result by a less elementary, but much more general argument, which was equally applicable to multi-dimensional problems. The following two theorems, due to Weyl, replace the problem of showing that a sequence is uniformly distributed mod 1 by a more tractable analytic problem.

Theorem 1 *A real sequence (ξ_n) is uniformly distributed mod 1 if and only if, for every function $f : I \to \mathbb{C}$ which is Riemann integrable,*

$$N^{-1}\sum_{n=1}^{N} f(\{\xi_n\}) \to \int_I f(t)\,dt \quad \text{as } N \to \infty. \tag{1}$$

Proof For any $\alpha, \beta \in I$ with $\alpha < \beta$, let $\chi_{\alpha,\beta}$ denote the *indicator function* of the interval $[\alpha, \beta)$, i.e.

$$\begin{aligned} \chi_{\alpha,\beta}(t) &= 1 \quad \text{for } \alpha \le t < \beta, \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Since

$$\int_I \chi_{\alpha,\beta}(t)\,dt = \beta - \alpha,$$

the definition of uniform distribution can be rephrased by saying that the sequence (ξ_n) is uniformly distributed mod 1 if and only if, for all choices of α and β,

$$N^{-1}\sum_{n=1}^{N} \chi_{\alpha,\beta}(\{\xi_n\}) \to \int_I \chi_{\alpha,\beta}(t)\,dt \quad \text{as } N \to \infty.$$

Thus the sequence (ξ_n) is certainly uniformly distributed mod 1 if (1) holds for every Riemann integrable function f.

Suppose now that the sequence (ξ_n) is uniformly distributed mod 1. Then (1) holds not only for every function $f = \chi_{\alpha,\beta}$, but also for every finite linear combination of such functions, i.e. for every *step-function* f. But, for any real-valued Riemann integrable function f and any $\varepsilon > 0$, there exist step-functions f_1, f_2 such that

$$f_1(t) \le f(t) \le f_2(t) \quad \text{for every } t \in I$$

and

$$\int_I (f_2(t) - f_1(t))\,dt < \varepsilon.$$

Hence

$$\begin{aligned} N^{-1}\sum_{n=1}^{N} f(\{\xi_n\}) - \int_I f(t)\,dt &\le N^{-1}\sum_{n=1}^{N} f_2(\{\xi_n\}) - \int_I f_2(t)\,dt + \varepsilon \\ &< 2\varepsilon \quad \text{for all large } N, \end{aligned}$$

and similarly

$$N^{-1}\sum_{n=1}^{N} f(\{\xi_n\}) - \int_I f(t)\,dt > -2\varepsilon \quad \text{for all large } N.$$

Thus (1) holds when the Riemann integrable function f is real-valued and also, by linearity, when it is complex-valued. □

A converse of Theorem 1 has been proved by de Bruijn and Post (1968): if a function $f : I \to \mathbb{C}$ has the property that

$$\lim_{N\to\infty} N^{-1} \sum_{n=1}^{N} f(\{\xi_n\})$$

exists for every sequence (ξ_n) which is uniformly distributed mod 1, then f is Riemann integrable.

In the statement of the next result, and throughout the rest of the chapter, we use the abbreviation

$$e(t) = e^{2\pi i t}.$$

In the proof of the next result we use the *Weierstrass approximation theorem*: any continuous function $f : I \to \mathbb{C}$ of period 1 is the uniform limit of a sequence (f_n) of trigonometric polynomials. In fact, as Fejér (1904) showed, one can take f_n to be the arithmetic mean $(S_0 + \cdots + S_{n-1})/n$, where

$$S_m = S_m(x) := \sum_{h=-m}^{m} c_h e(hx)$$

is the m-th partial sum of the Fourier series for f. This yields the explicit formula

$$f_n(x) = \int_I K_n(x-t) f(t)\, dt,$$

where

$$K_n(u) = (\sin^2 n\pi u)/(n \sin^2 \pi u).$$

Theorem 2 *A real sequence (ξ_n) is uniformly distributed mod 1 if and only if, for every integer $h \neq 0$,*

$$N^{-1} \sum_{n=1}^{N} e(h\xi_n) \to 0 \quad \text{as } N \to \infty. \tag{2}$$

Proof If the sequence (ξ_n) is uniformly distributed mod 1 then, by taking $f(t) = e(ht)$ in Theorem 1 we obtain (2) since, for every integer $h \neq 0$,

$$\int_I e(ht)\, dt = 0.$$

Conversely, suppose (2) holds for every nonzero integer h. Then, by linearity, for any trigonometric polynomial

$$g(t) = \sum_{h=-m}^{m} b_h e(ht)$$

we have

$$N^{-1} \sum_{n=1}^{N} g(\{\xi_n\}) \to b_0 = \int_I g(t)\, dt \quad \text{as } N \to \infty.$$

If f is a continuous function of period 1 then, by the Weierstrass approximation theorem, for any $\varepsilon > 0$ there exists a trigonometric polynomial $g(t)$ such that $|f(t) - g(t)| < \varepsilon$ for every $t \in I$. Hence

$$\begin{aligned}
&\left| N^{-1} \sum_{n=1}^{N} f(\{\xi_n\}) - \int_I f(t)\,dt \right| \\
&\quad \le \left| N^{-1} \sum_{n=1}^{N} (f(\{\xi_n\}) - g(\{\xi_n\})) \right| + \left| N^{-1} \sum_{n=1}^{N} g(\{\xi_n\}) - \int_I g(t)\,dt \right| \\
&\qquad\qquad + \left| \int_I (g(t) - f(t))\,dt \right| \\
&\quad < 2\varepsilon + \left| N^{-1} \sum_{n=1}^{N} g(\{\xi_n\}) - \int_I g(t)\,dt \right| \\
&\quad < 3\varepsilon \quad \text{for all large } N.
\end{aligned}$$

Thus (1) holds for every continuous function f of period 1.

Finally, if $\chi_{\alpha,\beta}$ is the function defined in the proof of Theorem 1 then, for any $\varepsilon > 0$, there exist continuous functions f_1, f_2 of period 1 such that

$$f_1(t) \le \chi_{\alpha,\beta}(t) \le f_2(t) \quad \text{for every } t \in I$$

and

$$\int_I (f_2(t) - f_1(t))\,dt < \varepsilon,$$

from which it follows similarly that

$$N^{-1} \sum_{n=1}^{N} \chi_{\alpha,\beta}(\{\xi_n\}) \to \int_I \chi_{\alpha,\beta}(t)\,dt \quad \text{as } N \to \infty.$$

Thus the sequence (ξ_n) is uniformly distributed mod 1. □

Weyl's criterion, as Theorem 2 is usually called, immediately implies Bohl's result:

Proposition 3 *If ξ is irrational, the sequence $(n\xi)$ is uniformly distributed mod* 1.

Proof For any nonzero integer h,

$$e(h\xi) + e(2h\xi) + \cdots + e(Nh\xi) = (e((N+1)h\xi) - e(h\xi))/(e(h\xi) - 1).$$

Hence

$$\left| N^{-1} \sum_{n=1}^{N} e(hn\xi) \right| \le 2|e(h\xi) - 1|^{-1} N^{-1},$$

and the result follows from Theorem 2. □

These results can be immediately extended to higher dimensions. A sequence (x_n) of vectors in $\mathbb{R}^d$ is said to be *uniformly distributed mod* 1 if, for all vectors $a = (\alpha_1, \dots, \alpha_d)$ and $b = (\beta_1, \dots, \beta_d)$ with $0 \le \alpha_k < \beta_k \le 1$ $(k = 1, \dots, d)$,

$$\varphi_{a,b}(N)/N \to \prod_{k=1}^{d} (\beta_k - \alpha_k) \quad \text{as } N \to \infty,$$

where $x_n = (\xi_n^{(1)}, \dots, \xi_n^{(d)})$ and $\varphi_{a,b}(N)$ is the number of positive integers $n \le N$ such that $\alpha_k \le \{\xi_n^{(k)}\} < \beta_k$ for every $k \in \{1, \dots, d\}$. Let I^d be the set of all $x = (\xi^{(1)}, \dots, \xi^{(d)})$ such that $0 \le \xi^{(k)} \le 1$ $(k = 1, \dots, d)$ and, for an arbitrary vector $x = (\xi^{(1)}, \dots, \xi^{(d)})$, put

$$\{x\} = (\{\xi^{(1)}\}, \dots, \{\xi^{(d)}\}).$$

Then Theorems 1 and 2 have the following generalizations:

Theorem 1′ *A sequence (x_n) of vectors in $\mathbb{R}^d$ is uniformly distributed mod* 1 *if and only if, for every function $f : I^d \to \mathbb{C}$ which is Riemann integrable,*

$$N^{-1} \sum_{n=1}^{N} f(\{x_n\}) \to \int_I \cdots \int_I f(t_1, \dots, t_d)\, dt_1 \cdots dt_d \quad \text{as } N \to \infty.$$

Theorem 2′ *A sequence (x_n) of vectors in $\mathbb{R}^d$ is uniformly distributed mod* 1 *if and only if, for every nonzero vector $m = (\mu_1, \dots, \mu_d) \in \mathbb{Z}^d$,*

$$N^{-1} \sum_{n=1}^{N} e(m \cdot x_n) \to 0 \quad \text{as } N \to \infty,$$

where $m \cdot x_n = \mu_1 \xi_n^{(1)} + \cdots + \mu_d \xi_n^{(d)}$.

Proposition 3 can also be generalized in the following way:

Proposition 3′ *If $x = (\xi^{(1)}, \dots, \xi^{(d)})$ is any vector in $\mathbb{R}^d$ such that $1, \xi^{(1)}, \dots, \xi^{(d)}$ are linearly independent over the field $\mathbb{Q}$ of rational numbers, then the sequence (nx) is uniformly distributed mod* 1.

In particular, the sequence $(\{nx\}) = (\{n\xi^{(1)}\}, \dots, \{n\xi^{(d)}\})$ is dense in the d-dimensional unit cube if $1, \xi^{(1)}, \dots, \xi^{(d)}$ are linearly independent over the field $\mathbb{Q}$ of rational numbers. This much weaker assertion had already been proved before Weyl by Kronecker (1884).

It is easily seen that the linear independence of $1, \xi^{(1)}, \dots, \xi^{(d)}$ over the field $\mathbb{Q}$ of rational numbers is also necessary for the sequence $(\{nx\})$ to be dense in the d-dimensional unit cube and, *a fortiori*, for the sequence (nx) to be uniformly distributed mod 1. For if $1, \xi^{(1)}, \dots, \xi^{(d)}$ are linearly dependent over $\mathbb{Q}$ there exists a nonzero vector $m = (\mu_1, \dots, \mu_d) \in \mathbb{Z}^d$ such that

$$m \cdot x = \mu_1 \xi^{(1)} + \cdots + \mu_d \xi^{(d)} \in \mathbb{Z}.$$

It follows that each point of the sequence (nx) lies on some hyperplane $m \cdot y = h$, where $h \in \mathbb{Z}$. Without loss of generality, suppose $\mu_1 \neq 0$. Then no point of the d-dimensional unit cube which is sufficiently close to the point $(|2\mu_1|^{-1}, 0, \ldots, 0)$ lies on such a hyperplane.

We now return to the one-dimensional case. Weyl used Theorem 2 to prove, not only Proposition 3, but also a deeper result concerning the uniform distribution of the sequence $(f(n))$, where f is a polynomial of any positive degree. We will derive Weyl's result by a more general argument due to van der Corput (1931), based on the following inequality:

Lemma 4 *If $\zeta_1, \ldots, \zeta_N$ are arbitrary complex numbers then, for any positive integer $M \leq N$,*

$$M^2\left|\sum_{n=1}^{N}\zeta_n\right|^2 \leq M(M+N-1)\sum_{n=1}^{N}|\zeta_n|^2 + 2(M+N-1)\sum_{m=1}^{M-1}(M-m)\left|\sum_{n=1}^{N-m}\overline{\zeta_n}\zeta_{n+m}\right|.$$

Proof Put $\zeta_n = 0$ if $n \leq 0$ or $n > N$. Then it is easily verified that

$$M\sum_{n=1}^{N}\zeta_n = \sum_{h=1}^{M+N-1}\left(\sum_{k=0}^{M-1}\zeta_{h-k}\right).$$

Applying Schwarz's inequality (Chapter I, §4), we get

$$\begin{aligned} M^2\left|\sum_{n=1}^{N}\zeta_n\right|^2 &\leq (M+N-1)\sum_{h=1}^{M+N-1}\left|\sum_{k=0}^{M-1}\zeta_{h-k}\right|^2 \\ &= (M+N-1)\sum_{h=1}^{M+N-1}\sum_{j,k=0}^{M-1}\zeta_{h-k}\overline{\zeta_{h-j}}. \end{aligned}$$

On the right side any term $|\zeta_n|^2$ occurs exactly M times, namely for $h-k = h-j = n$. A term $\overline{\zeta_n}\zeta_{n+m}$ or $\zeta_n\overline{\zeta_{n+m}}$, where $m > 0$, occurs only if $m < M$ and then it occurs exactly $M-m$ times. Thus the right side is equal to

$$M(M+N-1)\sum_{n=1}^{N}|\zeta_n|^2 + (M+N-1)\sum_{m=1}^{M-1}(M-m)\sum_{n=1}^{N-m}(\overline{\zeta_n}\zeta_{n+m} + \zeta_n\overline{\zeta_{n+m}}).$$

The lemma follows. □

Corollary 5 *If (ξ_n) is a real sequence such that, for each positive integer m,*

$$N^{-1}\sum_{n=1}^{N}e(\xi_{n+m} - \xi_n) \to 0 \quad \text{as } N \to \infty,$$

then

$$N^{-1}\sum_{n=1}^{N}e(\xi_n) \to 0 \quad \text{as } N \to \infty.$$

Proof By taking $\zeta_n = e(\xi_n)$ in Lemma 4 we obtain, for $1 \leq M \leq N$,

$$N^{-2}\left|\sum_{n=1}^{N} e(\xi_n)\right|^2 \leq 2(M+N-1)M^{-2}N^{-2}\sum_{m=1}^{M-1}(M-m)\left|\sum_{n=1}^{N-m} e(\xi_{n+m}-\xi_n)\right| + (M+N-1)M^{-1}N^{-1}.$$

Keeping M fixed and letting $N \to \infty$, we get

$$\overline{\lim_{N\to\infty}}\, N^{-2}\left|\sum_{n=1}^{N} e(\xi_n)\right|^2 \leq M^{-1}.$$

But M can be chosen as large as we please. □

An immediate consequence is van der Corput's *difference theorem*:

Proposition 6 *The real sequence* (ξ_n) *is uniformly distributed mod* 1 *if, for each positive integer* m*, the sequence* $(\xi_{n+m} - \xi_n)$ *is uniformly distributed mod* 1.

Proof If the sequences $(\xi_{n+m} - \xi_n)$ are uniformly distributed mod 1 then, by Theorem 2,

$$N^{-1}\sum_{n=1}^{N} e(h(\xi_{n+m}-\xi_n)) \to 0 \quad \text{as } N \to \infty$$

for all integers $h \neq 0, m > 0$. Replacing ξ_n by $h\xi_n$ in Corollary 5 we obtain, for all integers $h \neq 0$,

$$N^{-1}\sum_{n=1}^{N} e(h\xi_n) \to 0 \quad \text{as } N \to \infty.$$

Hence, by Theorem 2 again, the sequence (ξ_n) is uniformly distributed mod 1. □

The sequence $(n\xi)$, with ξ irrational, shows that we cannot replace 'if' by 'if and only if' in the statement of Proposition 6. Weyl's result will now be derived from Proposition 6:

Proposition 7 *If*

$$f(t) = \alpha_r t^r + \alpha_{r-1}t^{r-1} + \cdots + \alpha_0$$

is any polynomial with real coefficients α_k *such that* α_k *is irrational for at least one* $k > 0$*, then the sequence* $(f(n))$ *is uniformly distributed mod* 1.

Proof If $r = 1$, then the result holds by the same argument as in Proposition 3. We assume that $r > 1$, $\alpha_r \neq 0$ and the result holds for polynomials of degree less than r.

For any positive integer m,

$$g_m(t) = f(t+m) - f(t)$$

is a polynomial of degree $r-1$ with leading coefficient $rm\alpha_r$. If α_r is irrational, then $rm\alpha_r$ is also irrational and hence, by the induction hypothesis, the sequence $(g_m(n))$ is uniformly distributed mod 1. Consequently, by Proposition 6, the sequence $(f(n))$ is also uniformly distributed mod 1.

Suppose next that the leading coefficient α_r is rational, and let $\alpha_s (1 \le s < r)$ be the coefficient nearest to it which is irrational. Then the coefficients of $t^{r-1}, \dots, t^s$ of the polynomial $g_m(t)$ are rational, but the coefficient of t^{s-1} is irrational. If $s > 1$, it follows again from the induction hypothesis and Proposition 6 that the sequence $(f(n))$ is uniformly distributed mod 1.

Suppose finally that $s = 1$ and put

$$F(t) = \alpha_r t^r + \alpha_{r-1} t^{r-1} + \cdots + \alpha_2 t^2.$$

If $q > 0$ is a common denominator for the rational numbers $\alpha_2, \dots, \alpha_r$ then, for any integer $h \neq 0$ and any nonnegative integers j, k,

$$e(hF(jq+k)) = e(hF(k)).$$

Write $N = \ell q + k$, where $\ell = \lfloor N/q \rfloor$ and $0 \le k < q$. Since $f(t) = F(t) + \alpha_1 t + \alpha_0$, we obtain

$$\begin{aligned} N^{-1} \sum_{n=0}^{N-1} e(hf(n)) &= N^{-1} \sum_{k=0}^{q-1} \sum_{j=0}^{\ell-1} e(hf(jq+k)) + N^{-1} \sum_{n=\ell q}^{N} e(hf(n)) \\ &= N^{-1} \lfloor N/q \rfloor \sum_{k=0}^{q-1} e(hF(k)) \sum_{j=0}^{\ell-1} \ell^{-1} e(h(jq\alpha_1 + k\alpha_1 + \alpha_0)) \\ &\quad + N^{-1} \sum_{n=\ell q}^{N} e(hf(n)). \end{aligned}$$

The last term tends to zero as $N \to \infty$, since the sum contains at most q terms, each of absolute value 1. By Theorem 2, each of the q inner sums in the first term also tends to zero as $N \to \infty$, because the result holds for $r = 1$. Hence, by Theorem 2 again, the sequence $(f(n))$ is uniformly distributed mod 1. □

An interesting extension of Proposition 6 was derived by Korobov and Postnikov (1952):

Proposition 8 *If, for every positive integer m, the sequence $(\xi_{n+m} - \xi_n)$ is uniformly distributed mod* 1 *then, for all integers $q > 0$ and $r \ge 0$, the sequence (ξ_{qn+r}) is uniformly distributed mod* 1.

Proof We may suppose $q > 1$, since the assertion follows at once from Proposition 6 if $q = 1$. By Theorem 2 it is enough to show that, for every integer $m \neq 0$,

$$S := N^{-1} \sum_{n=1}^{N} e(m\xi_{qn+r}) \to 0 \quad \text{as } N \to \infty.$$

Since

$$q^{-1}\sum_{k=1}^{q} e(nk/q) = 1 \quad \text{if } n \equiv 0 \bmod q,$$
$$= 0 \quad \text{if } n \not\equiv 0 \bmod q,$$

we can write

$$S = (qN)^{-1}\sum_{n=1}^{qN} e(m\xi_{n+r})\sum_{k=1}^{q} e(nk/q)$$
$$= (qN)^{-1}\sum_{k=1}^{q}\sum_{n=1}^{qN} e(m\eta_n^{(k)}),$$

where we have put

$$\eta_n^{(k)} = \xi_{n+r} + nk/mq.$$

By hypothesis, for every positive integer h, the sequence

$$\eta_{n+h}^{(k)} - \eta_n^{(k)} = \xi_{n+h+r} - \xi_{n+r} - hk/mq$$

is uniformly distributed mod 1. Hence $\eta_n^{(k)}$ is uniformly distributed mod 1, by Proposition 6. Thus, for each $k \in \{1, \ldots, q\}$,

$$(qN)^{-1}\sum_{n=1}^{qN} e(m\eta_n^{(k)}) \to 0 \quad \text{as } N \to \infty,$$

and consequently also $S \to 0$ as $N \to \infty$. □

As an application of Proposition 8 we prove

Proposition 9 *Let A be a $d \times d$ matrix of integers, no eigenvalue of which is a root of unity. If, for some $x \in \mathbb{R}^d$, the sequence $(A^n x)$ is uniformly distributed mod* 1 *then, for any integers $q > 0$ and $r \geq 0$, the sequence $(A^{qn+r}x)$ is also uniformly distributed mod* 1.

Proof It follows from Theorem 2′ that, for any nonzero vector $m \in \mathbb{Z}^d$, the scalar sequence $\xi_n = m \cdot A^n x$ is uniformly distributed mod 1. For any positive integer h, the sequence

$$\xi_{n+h} - \xi_n = m \cdot (A^h - I)A^n x = (A^h - I)^t m \cdot A^n x$$

has the same form as the sequence ξ_n, since the hypotheses ensure that $(A^h - I)^t m$ is a nonzero vector in $\mathbb{Z}^d$. Hence the sequence $\xi_{n+h} - \xi_n$ is uniformly distributed mod 1. Therefore, by Proposition 8, the sequence $\xi_{qn+r} = m \cdot A^{qn+r}x$ is uniformly distributed mod 1, and thus the sequence $A^{qn+r}x$ is uniformly distributed mod 1. □

It may be noted that the matrix A in Proposition 9 is necessarily non-singular. For if $\det A = 0$, there exists a nonzero vector $z \in \mathbb{Z}^d$ such that $A^t z = 0$. Then, for any $x \in \mathbb{R}^d$ and any positive integer n, $e(z \cdot A^n x) = e((A^t)^n z \cdot x) = 1$. Thus $N^{-1} \sum_{n=1}^{N} e(z \cdot A^n x) = 1$ and therefore, by Theorem 2′, the sequence $A^n x$ is not uniformly distributed mod 1.

Further examples of uniformly distributed sequences are provided by the following result, which is due to Fejér (c. 1924):

Proposition 10 *Let (ξ_n) be a sequence of real numbers such that $\eta_n := \xi_{n+1} - \xi_n$ tends to zero monotonically as $n \to \infty$. Then (ξ_n) is uniformly distributed mod 1 if $n|\eta_n| \to \infty$ as $n \to \infty$.*

Proof By changing the signs of all ξ_n we may restrict attention to the case where the sequence (η_n) is strictly decreasing. For any real numbers α, β we have

$$\begin{aligned} |e(\alpha) - e(\beta) - 2\pi i(\alpha - \beta)e(\beta)| &= |e(\alpha - \beta) - 1 - 2\pi i(\alpha - \beta)| \\ &= 4\pi^2 \left| \int_0^{\alpha-\beta} (\alpha - \beta - t)e(t)\, dt \right| \\ &\le 4\pi^2 \left| \int_0^{\alpha-\beta} (\alpha - \beta - t)\, dt \right| \\ &= 2\pi^2(\alpha - \beta)^2. \end{aligned}$$

If we take $\alpha = h\xi_{n+1}$ and $\beta = h\xi_n$, where h is any nonzero integer, this yields

$$|e(h\xi_{n+1})/\eta_n - e(h\xi_n)/\eta_n - 2\pi i h e(h\xi_n)| \le 2\pi^2 h^2 \eta_n$$

and hence

$$|e(h\xi_{n+1})/\eta_{n+1} - e(h\xi_n)/\eta_n - 2\pi i h e(h\xi_n)| \le 1/\eta_{n+1} - 1/\eta_n + 2\pi^2 h^2 \eta_n.$$

Taking $n = 1, \ldots, N$ and adding, we obtain

$$\begin{aligned} \left| 2\pi h \sum_{n=1}^{N} e(h\xi_n) \right| &\le 1/\eta_{N+1} + 1/\eta_1 + \sum_{n=1}^{N} (1/\eta_{n+1} - 1/\eta_n) + 2\pi^2 h^2 \sum_{n=1}^{N} \eta_n \\ &= 2/\eta_{N+1} + 2\pi^2 h^2 \sum_{n=1}^{N} \eta_n. \end{aligned}$$

Thus

$$N^{-1} \left| \sum_{n=1}^{N} e(h\xi_n) \right| \le (\pi |h| N \eta_{N+1})^{-1} + \pi |h| N^{-1} \sum_{n=1}^{N} \eta_n.$$

But the right side of this inequality tends to zero as $N \to \infty$, since $N\eta_N \to \infty$ and $\eta_N \to 0$. □

By the mean value theorem, the hypotheses of Proposition 10 are certainly satisfied if $\xi_n = f(n)$, where f is a differentiable function such that $f'(t) \to 0$ monotonically as $t \to \infty$ and $t|f'(t)| \to \infty$ as $t \to \infty$. Consequently the sequence (an^α) is uniformly distributed mod 1 if $a \neq 0$ and $0 < \alpha < 1$, and the sequence $(a(\log n)^\alpha)$ is uniformly distributed mod 1 if $a \neq 0$ and $\alpha > 1$. By using van der Corput's difference theorem and an inductive argument starting from Proposition 10, it may be further shown that the sequence (an^α) is uniformly distributed mod 1 for any $a \neq 0$ and any $\alpha > 0$ which is not an integer.

It has been shown by Kemperman (1973) that 'if' may be replaced by 'if and only if' in the statement of Proposition 10. Consequently the sequence $(a(\log n)^\alpha)$ is not uniformly distributed mod 1 if $0 < \alpha \leq 1$.

The theory of uniform distribution has an application, and its origin, in astronomy. In his investigations on the secular perturbations of planetary orbits Lagrange (1782) was led to the problem of *mean motion*: if

$$z(t) = \sum_{k=1}^{n} \rho_k e(\omega_k t + \alpha_k),$$

where $\rho_k > 0$ and $\alpha_k, \omega_k \in \mathbb{R}$ $(k = 1, \ldots, n)$, does $t^{-1} \arg z(t)$ have a finite limit as $t \to +\infty$? It is assumed that $z(t)$ never vanishes and $\arg z(t)$ is then defined by continuity. (Zeros of $z(t)$ can be admitted by writing $z(t) = \rho(t)e(\phi(t))$, where $\rho(t)$ and $\phi(t)$ are continuous real-valued functions and $\rho(t)$ is required to change sign at a zero of $z(t)$ of odd multiplicity.)

In the astronomical application $\arg z(t)$ measures the longitude of the perihelion of the planetary orbit. Lagrange showed that the limit

$$\mu = \lim_{t \to +\infty} t^{-1} \arg z(t)$$

does exist when $n = 2$ and also, for arbitrary n, when some ρ_k exceeds the sum of all the others. The only planets which do not satisfy this second condition are Venus and Earth. Lagrange went on to say that, when neither of the two conditions was satisfied, the problem was "very difficult and perhaps impossible".

There was no further progress until the work of Bohl (1909), who took $n = 3$ and considered the non-Lagrangian case when there exists a triangle with sidelengths ρ_1, ρ_2, ρ_3. He showed that the limit μ exists if $\omega_1, \omega_2, \omega_3$ are linearly independent over the rational field $\mathbb{Q}$ and then $\mu = \lambda_1\omega_1 + \lambda_2\omega_2 + \lambda_3\omega_3$, where $\pi\lambda_1, \pi\lambda_2, \pi\lambda_3$ are the angles of the triangle with sidelengths ρ_1, ρ_2, ρ_3. In the course of the proof he stated and proved Proposition 3 (without formulating the general concept of uniform distribution).

Using his earlier results on uniform distribution, Weyl (1938) showed that the limit μ exists if $\omega_1, \ldots, \omega_n$ are linearly independent over the rational field $\mathbb{Q}$ and then

$$\mu = \lambda_1\omega_1 + \cdots + \lambda_n\omega_n,$$

where $\lambda_k \geq 0$ $(k = 1, \ldots, n)$ and $\sum_{k=1}^{n} \lambda_k = 1$. The coefficients λ_k depend only on the ρ's, not on the α's or ω's, and there is even an explicit expression for λ_k, involving Bessel functions, which is derived from the theory of random walks.

Finally, it was shown by Jessen and Tornehave (1945) that the limit μ exists for arbitrary $\omega_k \in \mathbb{R}$.

2 Discrepancy

The *star discrepancy* of a finite set of points $\xi_1, \ldots, \xi_N$ in the unit interval $I = [0, 1]$ is defined to be

$$D_N^* = D_N^*(\xi_1, \ldots, \xi_N) = \sup_{0<\alpha\le 1} |\varphi_\alpha(N)/N - \alpha|,$$

where $\varphi_\alpha(N) = \varphi_{0,\alpha}(N)$ denotes the number of positive integers $n \le N$ such that $0 \le \xi_n < \alpha$. Here we will omit the qualifier 'star', since we will not be concerned with any other type of discrepancy and the notation D_N^* should provide adequate warning.

It was discovered only in 1972, by Niederreiter, that the preceding definition may be reformulated in the following simple way:

Proposition 11 *If $\xi_1, \ldots, \xi_N$ are real numbers such that $0 \le \xi_1 \le \cdots \le \xi_N \le 1$, then*

$$\begin{aligned} D_N^* = D_N^*(\xi_1, \ldots, \xi_N) &= \max_{1\le k\le N} \max(|\xi_k - k/N|, |\xi_k - (k-1)/N|) \\ &= (2N)^{-1} + \max_{1\le k\le N} |\xi_k - (2k-1)/2N|. \end{aligned}$$

Proof Put $\xi_0 = 0, \xi_{N+1} = 1$. Since the distinct ξ_k with $0 \le k \le N+1$ define a subdivision of the unit interval I, we have

$$\begin{aligned} D_N^* &= \max_{k:\xi_k<\xi_{k+1}} \sup_{\xi_k\le\alpha<\xi_{k+1}} |\varphi_\alpha(N)/N - \alpha| \\ &= \max_{k:\xi_k<\xi_{k+1}} \sup_{\xi_k\le\alpha<\xi_{k+1}} |k/N - \alpha|. \end{aligned}$$

But the function $f_k(t) = |k/N - t|$ attains its maximum in the interval $\xi_k \le t \le \xi_{k+1}$ at one of the endpoints of this interval. Consequently

$$D_N^* = \max_{k:\xi_k<\xi_{k+1}} \max(|k/N - \xi_k|, |k/N - \xi_{k+1}|).$$

We are going to show that in fact

$$D_N^* = \max_{0\le k\le N} \max(|k/N - \xi_k|, |k/N - \xi_{k+1}|).$$

Suppose $\xi_k < \xi_{k+1} = \xi_{k+2} = \cdots = \xi_{k+r} < \xi_{k+r+1}$ for some $r \ge 2$. By applying the same reasoning as before to the function $g_k(t) = |t - \xi_{k+1}|$ we obtain, for $1 \le j < r$,

$$\begin{aligned} |(k+j)/N - \xi_{k+j}| &= |(k+j)/N - \xi_{k+j+1}| = |(k+j)/N - \xi_{k+1}| \\ &< \max(|k/N - \xi_{k+1}|, |(k+r)/N - \xi_{k+1}|) \\ &= \max(|k/N - \xi_{k+1}|, |(k+r)/N - \xi_{k+r}|). \end{aligned}$$

Since both terms in the last maximum appear in the expression already obtained for D_N^*, it follows that this expression is not altered by dropping the restriction to those k for which $\xi_k < \xi_{k+1}$.

Since $|0/N - \xi_0| = |N/N - \xi_{N+1}| = 0$, we can now also write

$$D_N^* = \max_{1\le k\le N} \max(|k/N - \xi_k|, |(k-1)/N - \xi_k|).$$

The second expression for D_N^* follows immediately, since

$$\max(|k/N - \alpha|, |(k-1)/N - \alpha|) = |(k-1/2)/N - \alpha| + 1/2N. \qquad \square$$

Corollary 12 *If $\xi_1, \dots, \xi_N$ are real numbers such that $0 \le \xi_1 \le \cdots \le \xi_N \le 1$, then $D_N^* \ge (2N)^{-1}$. Moreover, equality holds if and only if $\xi_k = (2k-1)/N$ for $k = 1, \dots, N$.*

Thus Proposition 11 says that the discrepancy of any set of N points of I is obtained by adding to its minimal value $1/2N$ the maximum deviation of the set from the unique minimizing set, when both sets are arranged in order of magnitude.

The next result shows that the discrepancy $D_N^*(\xi_1, \dots, \xi_N)$ is a continuous function of $\xi_1, \dots, \xi_N$.

Proposition 13 *If $\xi_1, \dots, \xi_N$ and $\eta_1, \dots, \eta_N$ are two sets of N points of I, with the discrepancies D_N^* and E_N^* respectively, then*

$$|D_N^* - E_N^*| \le \max_{1\le k\le N} |\xi_k - \eta_k|.$$

Proof Let $x_1 \le \cdots \le x_N$ and $y_1 \le \cdots \le y_N$ be the two given sets rearranged in order of magnitude. It is enough to show that

$$\max_{1\le k\le N} |x_k - y_k| \le \delta := \max_{1\le k\le N} |\xi_k - \eta_k|,$$

since it then follows from Proposition 11 that

$$D_N^* \le \delta + E_N^*, \quad E_N^* \le \delta + D_N^*.$$

Assume, on the contrary, that $|x_k - y_k| > \delta$ for some k. Then either $x_k > y_k + \delta$ or $y_k > x_k + \delta$. Without loss of generality we restrict attention to the first case. By hypothesis, for each y_i with $1 \le i \le k$ there exists an x_{j_i} with $1 \le j_i \le N$ such that $|y_i - x_{j_i}| \le \delta$ and such that the subscripts j_i are distinct. Since $y_1 \le \cdots \le y_k$, it follows that

$$x_{j_i} \le y_i + \delta \le y_k + \delta < x_k.$$

But this is a contradiction, since there are at most $k-1$ x's less than x_k. $\square$

We now show how the notion of discrepancy makes it possible to obtain estimates for the accuracy of various methods of numerical integration.

Proposition 14 *If the function f satisfies the 'Lipschitz condition'*

$$|f(t_2) - f(t_1)| \le L|t_2 - t_1| \quad \text{for all } t_1, t_2 \in I,$$

then for any finite set $\xi_1, \dots, \xi_N \in I$ with discrepancy D_N^*,

$$\left| N^{-1} \sum_{n=1}^{N} f(\xi_n) - \int_I f(t)\,dt \right| \le L D_N^*.$$

Proof Without loss of generality we may assume $\xi_1 \le \cdots \le \xi_N$. Writing

$$\int_I f(t)\,dt = \sum_{n=1}^{N} \int_{(n-1)/N}^{n/N} f(t)\,dt,$$

we obtain

$$\begin{aligned}\left| N^{-1} \sum_{n=1}^{N} f(\xi_n) - \int_I f(t)\,dt \right| &\le \sum_{n=1}^{N} \int_{(n-1)/N}^{n/N} |f(\xi_n) - f(t)|\,dt \\ &\le L \sum_{n=1}^{N} \int_{(n-1)/N}^{n/N} |\xi_n - t|\,dt.\end{aligned}$$

But for $(n-1)/N \le t \le n/N$ we have

$$|\xi_n - t| \le \max(|\xi_n - n/N|, |\xi_n - (n-1)/N|) \le D_N^*,$$

by Proposition 11. The result follows. □

As Koksma (1942) first showed, Proposition 14 can be sharpened in the following way:

Proposition 15 *If the function f has bounded variation on the unit interval I, with total variation V, then for any finite set $\xi_1, \ldots, \xi_N \in I$ with discrepancy D_N^*,*

$$\left| N^{-1} \sum_{n=1}^{N} f(\xi_n) - \int_I f(t)\,dt \right| \le V D_N^*.$$

Proof Without loss of generality we may assume $\xi_1 \le \cdots \le \xi_N$ and we put $\xi_0 = 0$, $\xi_{N+1} = 1$. By integration and summation by parts we obtain

$$\begin{aligned}\sum_{n=0}^{N} \int_{\xi_n}^{\xi_{n+1}} (t - n/N)\,df(t) &= \int_I t\,df(t) - N^{-1} \sum_{n=0}^{N} n(f(\xi_{n+1}) - f(\xi_n)) \\ &= [tf(t)]_0^1 - \int_I f(t)\,dt - f(1) + N^{-1} \sum_{n=0}^{N-1} f(\xi_{n+1}) \\ &= N^{-1} \sum_{n=1}^{N} f(\xi_n) - \int_I f(t)\,dt.\end{aligned}$$

The result follows, since for $\xi_n \le t \le \xi_{n+1}$ we have

$$|t - n/N| \le \max(|\xi_n - n/N|, |\xi_{n+1} - n/N|) \le D_N^*.$$ □

As an application of Proposition 15 we prove

Proposition 16 *If $\xi_1, \dots, \xi_N$ are points of the unit interval I with discrepancy D_N^* then, for any integer $h \neq 0$,*

$$\left| N^{-1} \sum_{n=1}^{N} e(h\xi_n) \right| \leq 4|h| D_N^*.$$

Proof We can write

$$N^{-1} \sum_{n=1}^{N} e(h\xi_n) = \rho e(\alpha),$$

where $\rho \geq 0$ and $\alpha \in I$. Thus

$$\rho = N^{-1} \sum_{n=1}^{N} e(h\xi_n - \alpha).$$

Adding this relation to its complex conjugate, we obtain

$$\rho = N^{-1} \sum_{n=1}^{N} \cos 2\pi (h\xi_n - \alpha).$$

The result follows by applying Proposition 15 to the function $f(t) = \cos 2\pi (ht - \alpha)$, which has bounded variation on I with total variation $\int_I |f'(t)|\, dt = 4|h|$. □

An inequality in the opposite direction to Proposition 16 was obtained by Erdős and Turan (1948) who showed that, for any positive integer m,

$$D_N^* \leq C\left(m^{-1} + \sum_{h=1}^{m} h^{-1} \left| N^{-1} \sum_{n=1}^{N} e(h\xi_n) \right| \right),$$

where the positive constant C is independent of m, N and the ξ's. Niederreiter and Philipp (1973) showed that one can take $C = 4$. Furthermore they generalized the result and simplified the proof.

The connection between these results and the theory of uniform distribution is close at hand. Let (ξ_n) be an arbitrary sequence of real numbers and let δ_N denote the discrepancy of the fractional parts $\{\xi_1\}, \dots, \{\xi_N\}$. By the remark after the definition of uniform distribution in §1, *the sequence* (ξ_n) *is uniformly distributed mod* 1 *if and only if* $\delta_N \to 0$ as $N \to \infty$. It follows from Proposition 16 and the inequality of Erdős and Turan (in which m may be arbitrarily large) that $\delta_N \to 0$ as $N \to \infty$ if and only if, for every integer $h \neq 0$,

$$N^{-1} \sum_{n=1}^{N} e(h\xi_n) \to 0 \quad \text{as } N \to \infty.$$

This provides a new proof of Theorem 2. Furthermore, from bounds for the exponential sums we can obtain estimates for the rapidity with which δ_N tends to zero.

Propositions 14 and 15 show that in a formula for numerical integration the nodes (ξ_n) should be chosen to have as small a discrepancy as possible. For a given finite number N of nodes Corollary 12 shows how this can be achieved. In practice, however, one does not know in advance an appropriate choice of N, since universal error bounds may grossly overestimate the error in a specific case. Consequently it is also of interest to consider the problem of choosing an infinite sequence (ξ_n) of nodes so that the discrepancy δ_N of $\xi_1, \ldots, \xi_N$ tends to zero as rapidly as possible when $N \to \infty$. There is a limit to what can be achieved in this way. W. Schmidt (1972), improving earlier results of van Aardenne-Ehrenfest (1949) and Roth (1954), showed that there exists an absolute constant $C > 0$ such that

$$\overline{\lim_{N\to\infty}}\, N\delta_N / \log N \geq C$$

for *every* infinite sequence (ξ_n). Kuipers and Niederreiter (1974) showed that a possible value for C was $(132 \log 2)^{-1} = 0.0109\ldots$, which Bejian (1979) improved to $(24 \log 2)^{-1} = 0.0601\ldots$.

Schmidt's result is best possible, apart from the value of the constant. Ostrowski (1922) had already shown that for the sequence $(\{n\alpha\})$, where $\alpha \in (0, 1)$ is irrational, one has

$$s^*(\alpha) := \overline{\lim_{N\to\infty}}\, N\delta_N / \log N < \infty$$

if in the continued fraction expansion

$$\alpha = [0; a_1, a_2, \ldots] = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cdots}}$$

the partial quotients a_k are bounded. Dupain and Sós (1984) have shown that the minimum value of $s^*(\alpha)$, for all such α, is $(4 \log(1+\sqrt{2}))^{-1} = 0.283\ldots$ and the minimum is attained for $\alpha = \sqrt{2} - 1 = [0; 2, 2, \ldots]$. Schoessengeier (1984) has proved that, for any irrational $\alpha \in (0, 1)$, one has $N\delta_N = O(\log N)$ if and only if the partial quotients a_k satisfy $\sum_{k=1}^{n} a_k = O(n)$.

There are other low discrepancy sequences. Haber (1966) showed that, for a sequence (ξ_n) constructed by van der Corput (1935),

$$\overline{\lim_{N\to\infty}}\, N\delta_N / \log N = (3 \log 2)^{-1} = 0.481\ldots.$$

van der Corput's sequence is defined as follows: if $n - 1 = a_m 2^m + \cdots + a_1 2^1 + a_0$, where $a_k \in \{0, 1\}$, then $\xi_n = a_0 2^{-1} + a_1 2^{-2} + \cdots + a_m 2^{-m-1}$. In other words, the expression for ξ_n in the base 2 is obtained from that for $n - 1$ by reflection in the 'decimal' point, a construction which is easily implemented on a computer. Various generalizations of this construction have been given, and Faure (1981) defined in this way a sequence (ξ_n) for which

$$\overline{\lim_{N\to\infty}}\, N\delta_N / \log N = (1919)(3454 \log 12)^{-1} = 0.223\ldots.$$

Thus if C^* is the least upper bound for all admissible values of C in Schmidt's result then, by what has been said, $0.060\ldots \le C^* \le 0.223\ldots$. It is natural to ask: what is the exact value of C^*, and is there a sequence (ξ_n) for which it is attained?

The notion of discrepancy is easily extended to higher dimensions by defining the discrepancy of a finite set of vectors $x_1, \ldots, x_N$ in the d-dimensional unit cube $I^d = I \times \cdots \times I$ to be

$$D_N^*(x_1, \ldots, x_N) = \sup_{0<\alpha_k\le 1 \ (k=1,\ldots,d)} |\varphi_a(N)/N - \alpha_1 \cdots \alpha_d|,$$

where $x_n = (\xi_n^{(1)}, \ldots, \xi_n^{(d)})$, $a = (\alpha_1, \ldots, \alpha_d)$ and $\varphi_a(N)$ is the number of positive integers $n \le N$ such that $0 \le \xi_n^{(k)} < \alpha_k$ for every $k \in \{1, \ldots, d\}$.

For $d > 1$ there is no simple reformulation of the definition analogous to Proposition 11, but many results do carry over. In particular, Proposition 15 was generalized and applied to the numerical evaluation of multiple integrals by Hlawka (1961/62). Indeed this application has greater value in higher dimensions, where other methods perform poorly.

For the application one requires a set of vectors $x_1, \ldots, x_N \in I^d$ whose discrepancy $D_N^*(x_1, \ldots, x_N)$ is small. A simple procedure for obtaining such a set, which is most useful when the integrand is smooth and has period 1 in each of its variables, is the method of 'good lattice points' introduced by Korobov (1959). Here, for a suitably chosen $g \in \mathbb{Z}^d$, one takes $x_n = \{(n-1)g/N\}$ $(n = 1, \ldots, N)$. A result of Niederreiter (1986) implies that, for every $d \ge 2$ and every $N \ge 2$, one can choose g so that

$$ND_N^* \le (1 + \log N)^d + d2^d.$$

The van der Corput sequence has also been generalized to any finite number of dimensions by Halton (1960). He defined an infinite sequence (x_n) of vectors in $\mathbb{R}^d$ for which

$$\overline{\lim_{N\to\infty}}\, N\delta_N/(\log N)^d < \infty.$$

It is conjectured that for each $d > 1$ (as for $d = 1$) there exists an absolute constant $C_d > 0$ such that

$$\overline{\lim_{N\to\infty}}\, N\delta_N/(\log N)^d \ge C_d$$

for every infinite sequence (x_n) of vectors in $\mathbb{R}^d$. However, the best known result remains that of Roth (1954), in which the exponent d is replaced by $d/2$.

3 Birkhoff's Ergodic Theorem

In statistical mechanics there is a procedure for calculating the physical properties of a system by simply averaging over all possible states of the system. To justify this procedure Boltzmann (1871) introduced what he later called the 'ergodic hypothesis'. In the formulation of Maxwell (1879) this says that "the system, if left to itself in its

actual state of motion, will, sooner or later, pass through every phase which is consistent with the equation of energy". The word *ergodic*, coined by Boltzmann (1884), was a composite of the Greek words for 'energy' and 'path'. It was recognized by Poincaré (1894) that it was too much to ask that a path pass through every state on the same energy surface as its initial state, and he suggested instead that it pass arbitrarily close to every such state. Moreover, he observed that it would still be necessary to exclude certain exceptional initial states.

A breakthrough came with the work of G.D. Birkhoff (1931), who showed that Lebesgue measure was the appropriate tool for treating the problem. He established a deep and general result which says that, apart from a set of initial states of measure zero, there is a definite limiting value for the proportion of time which a path spends in any given measurable subset B of an energy surface X. The proper formulation for the ergodic hypothesis was then that this limiting value should coincide with the ratio of the measure of B to that of X, i.e. that 'the paths through almost all initial states should be uniformly distributed over arbitrary measurable sets'. It was not difficult to deduce that this was the case if and only if 'any invariant measurable subset of X either had measure zero or had the same measure as X'.

Birkhoff proved his theorem in the framework of classical mechanics and for *flows* with continuous time. We will prove his theorem in the abstract setting of probability spaces and for *cascades* with discrete time. The abstract formulation makes possible other applications, for which continuous time is not appropriate.

Let $\mathscr{B}$ be a σ*-algebra* of subsets of a given set X, i.e. a nonempty family of subsets of X such that

(B1) the complement of any set in $\mathscr{B}$ is again a set in $\mathscr{B}$,
(B2) the union of any finite or countable collection of sets in $\mathscr{B}$ is again a set in $\mathscr{B}$.

It follows that $X \in \mathscr{B}$, since $B \in \mathscr{B}$ implies $B^c := X \backslash B \in \mathscr{B}$ and $X = B \cup B^c$. Hence also $\emptyset = X^c \in \mathscr{B}$. Furthermore, the intersection of any finite or countable collection of sets in $\mathscr{B}$ is again a set in $\mathscr{B}$, since $\bigcap_n B_n = X \backslash \big(\bigcup_n B_n^c \big)$. Hence if $A, B \in \mathscr{B}$, then

$$B \backslash A = B \cap A^c \in \mathscr{B}$$

and the *symmetric difference*

$$A \Delta B := (B \backslash A) \cup (A \backslash B) \in \mathscr{B}.$$

The family of all subsets of X is certainly a σ-algebra. Furthermore, the intersection of any collection of σ-algebras is again a σ-algebra. It follows that, for any family $\mathscr{A}$ of subsets of X, there is a σ-algebra $\sigma(\mathscr{A})$ which contains $\mathscr{A}$ and is contained in every σ-algebra which contains $\mathscr{A}$. We call $\sigma(\mathscr{A})$ the σ-algebra of subsets of X *generated* by $\mathscr{A}$.

Suppose $\mathscr{B}$ is a σ-algebra of subsets of X and a function $\mu : \mathscr{B} \to \mathbb{R}$ is defined such that

(Pr1) $\mu(B) \geq 0$ for every $B \in \mathscr{B}$,
(Pr2) $\mu(X) = 1$,
(Pr3) if (B_n) is a sequence of pairwise disjoint sets in $\mathscr{B}$, then $\mu\big(\bigcup_n B_n \big) = \sum_n \mu(B_n)$.

Then μ is said to be a *probability measure* and the triple $(X, \mathscr{B}, \mu)$ is said to be a *probability space*.

It is easily seen that the definition implies

(i) $\mu(\emptyset) = 0$,
(ii) $\mu(B^c) = 1 - \mu(B)$,
(iii) $\mu(A) \le \mu(B)$ if $A, B \in \mathscr{B}$ and $A \subseteq B$,
(iv) $\mu(B_n) \to \mu(B)$ if (B_n) is a sequence of sets in $\mathscr{B}$ such that $B_1 \supseteq B_2 \supseteq \cdots$ and $B = \bigcap_n B_n$.

If a property of points in a probability space $(X, \mathscr{B}, \mu)$ holds for all $x \in B$, where $B \in \mathscr{B}$ and $\mu(B) = 1$, then the property is said to hold for (μ-) *almost all* $x \in X$, or simply *almost everywhere* (a.e.).

A function $f : X \to \mathbb{R}$ is *measurable* if, for every $\alpha \in \mathbb{R}$, the set $\{x \in X : f(x) < \alpha\}$ is in $\mathscr{B}$. Let $f : X \to [0, \infty)$ be measurable and for any partition $\mathscr{P}$ of X into finitely many pairwise disjoint sets $B_1, \ldots, B_n \in \mathscr{B}$, put

$$L_{\mathscr{P}}(f) = \sum_{k=1}^{n} f_k \, \mu(B_k),$$

where $f_k = \inf\{f(x) : x \in B_k\}$. We say that f is *integrable* if

$$\int_X f \, \mathrm{d}\mu := \sup_{\mathscr{P}} L_{\mathscr{P}}(f) < \infty.$$

The set of all measurable functions $f : X \to \mathbb{R}$ such that $|f|$ is integrable is denoted by $L(X, \mathscr{B}, \mu)$.

A map $T : X \to X$ is said to be a *measure-preserving transformation* of the probability space $(X, \mathscr{B}, \mu)$ if, for every $B \in \mathscr{B}$, the set $T^{-1}B = \{x \in X : Tx \in B\}$ is again in $\mathscr{B}$ and $\mu(T^{-1}B) = \mu(B)$. This is equivalent to $\mu(TB) = \mu(B)$ for every $B \in \mathscr{B}$ if the measure-preserving transformation T is *invertible*, i.e. if T is bijective and $TB \in \mathscr{B}$ for every $B \in \mathscr{B}$. However, we do not wish to restrict attention to the invertible case. Several important examples of measure-preserving transformations of probability spaces will be given in the next section.

Birkhoff's ergodic theorem, which is also known as the 'individual' or 'pointwise' ergodic theorem, has the following statement:

Theorem 17 *Let T be a measure-preserving transformation of the probability space $(X, \mathscr{B}, \mu)$. If $f \in L(X, \mathscr{B}, \mu)$ then, for almost all $x \in X$, the limit*

$$f^*(x) = \lim_{n \to \infty} n^{-1} \sum_{k=0}^{n-1} f(T^k x)$$

exists and $f^*(Tx) = f^*(x)$. Moreover, $f^* \in L(X, \mathscr{B}, \mu)$ and $\int_X f^* \mathrm{d}\mu = \int_X f \, \mathrm{d}\mu$.

Proof It is sufficient to prove the theorem for nonnegative functions, since we can write $f = f_+ - f_-$, where

$$f_+(x) = \max\{f(x), 0\}, \quad f_-(x) = \max\{-f(x), 0\},$$

and $f_+, f_- \in L(X, \mathscr{B}, \mu)$.

Put

$$\bar{f}(x) = \overline{\lim_{n\to\infty}}\, n^{-1}\sum_{k=0}^{n-1} f(T^k x), \quad \underline{f}(x) = \underline{\lim_{n\to\infty}}\, n^{-1}\sum_{k=0}^{n-1} f(T^k x).$$

Then $\bar{f}$ and $\underline{f}$ are μ-measurable functions since, for any sequence (g_n),

$$\overline{\lim_{n\to\infty}}\, g_n(x) = \inf_m (\sup_{n\geq m} g_n(x)), \quad \underline{\lim_{n\to\infty}}\, g_n(x) = \sup_m (\inf_{n\geq m} g_n(x)).$$

Moreover $\bar{f}(x) = \bar{f}(Tx)$, $\underline{f}(x) = \underline{f}(Tx)$ for every $x \in X$, since

$$(n+1)^{-1}\sum_{k=0}^{n} f(T^k x) = (n+1)^{-1} f(x) + (1+1/n)^{-1} n^{-1}\sum_{k=0}^{n-1} f(T^{k+1}x).$$

It is sufficient to show that

$$\int_X \bar{f}\,\mathrm{d}\mu \leq \int_X f\,\mathrm{d}\mu \leq \int_X \underline{f}\,\mathrm{d}\mu.$$

For then, since $\underline{f} \leq \bar{f}$, it follows that $\bar{f}(x) = \underline{f}(x) = f^*(x)$ for μ-almost all $x \in X$ and

$$\int_X f^*\mathrm{d}\mu = \int_X f\,\mathrm{d}\mu.$$

Fix some $M > 0$ and define the 'cut-off' function $\bar{f}_M$ by

$$\bar{f}_M(x) = \min\{M, \bar{f}(x)\}.$$

Then $\bar{f}_M$ is bounded and $\bar{f}_M(Tx) = \bar{f}_M(x)$ for every $x \in X$. Fix also any $\varepsilon > 0$. By the definition of $\bar{f}(x)$, for each $x \in X$ there exists a positive integer n such that

$$\bar{f}_M(x) \leq n^{-1}\sum_{k=0}^{n-1} f(T^k x) + \varepsilon. \qquad (*)$$

Thus if F_n is the set of all $x \in X$ for which $(*)$ holds and if $E_n = \bigcup_{k=1}^{n} F_k$, then $E_1 \subseteq E_2 \subseteq \cdots$ and $X = \bigcup_{n\geq 1} E_n$. Since the sets E_n are μ-measurable, we can choose N so large that $\mu(E_N) > 1 - \varepsilon/M$.

Put

$$\tilde{f}(x) = f(x) \quad \text{if } x \in E_N,$$
$$= \max\{f(x), M\} \quad \text{if } x \notin E_N.$$

Also, let $\tau(x)$ be the least positive integer $n \leq N$ for which $(*)$ holds if $x \in E_N$, and let $\tau(x) = 1$ if $x \notin E_N$. Since $\bar{f}_M$ is T-invariant, $(*)$ implies

$$\sum_{k=0}^{n-1} \bar{f}_M(T^k x) \leq \sum_{k=0}^{n-1} f(T^k x) + n\varepsilon$$

and hence

$$\sum_{k=0}^{\tau(x)-1} \bar{f}_M(T^k x) \le \sum_{k=0}^{\tau(x)-1} \tilde{f}(T^k x) + \tau(x)\varepsilon.$$

To estimate the sum $\sum_{k=0}^{L-1} \bar{f}_M(T^k x)$ for any $L > N$, we partition it into blocks of the form

$$\sum_{k=0}^{\tau(y)-1} \bar{f}_M(T^k y)$$

and a remainder block. More precisely, define inductively

$$n_0(x) = 0, \quad n_k(x) = n_{k-1}(x) + \tau(T^{n_{k-1}} x) \quad (k = 1, 2, \ldots)$$

and define h by $n_h(x) < L \le n_{h+1}(x)$. Then

$$\sum_{k=0}^{n_1(x)-1} \bar{f}_M(T^k x) \le \sum_{k=0}^{n_1(x)-1} \tilde{f}(T^k x) + \tau(x)\varepsilon,$$

$$\sum_{k=n_1(x)}^{n_2(x)-1} \bar{f}_M(T^k x) \le \sum_{k=n_1(x)}^{n_2(x)-1} \tilde{f}(T^k x) + \tau(T^{n_1} x)\varepsilon,$$

$$\cdots$$

$$\sum_{k=n_{h-1}(x)}^{n_h(x)-1} \bar{f}_M(T^k x) \le \sum_{k=n_{h-1}(x)}^{n_h(x)-1} \tilde{f}(T^k x) + \tau(T^{n_{h-1}} x)\varepsilon.$$

Since $n_h(x) < L$, we obtain by addition

$$\sum_{k=0}^{n_h(x)-1} \bar{f}_M(T^k x) \le \sum_{k=0}^{n_h(x)-1} \tilde{f}(T^k x) + L\varepsilon.$$

On the other hand, since $L \le n_{h+1}(x) \le n_h(x) + N$, we have

$$\sum_{k=n_h(x)}^{L-1} \bar{f}_M(T^k x) \le NM.$$

Since $\tilde{f} \ge 0$, it follows that

$$\sum_{k=0}^{L-1} \bar{f}_M(T^k x) \le \sum_{k=0}^{L-1} \tilde{f}(T^k x) + L\varepsilon + NM.$$

Dividing by L and integrating over X, we obtain

$$\int_X \bar{f}_M \, \mathrm{d}\mu \le \int_X \tilde{f} \, \mathrm{d}\mu + \varepsilon + NM/L,$$

since the measure-preserving nature of T implies that, for any $g \in L(X, \mathscr{B}, \mu)$,

$$\int_X g(Tx)\,\mathrm{d}\mu(x) = \int_X g(x)\,\mathrm{d}\mu(x).$$

Since

$$\int_X \tilde{f}\,\mathrm{d}\mu \le \int_X f\,\mathrm{d}\mu + \int_{X\setminus E_N} M\,\mathrm{d}\mu \le \int_X f\,\mathrm{d}\mu + \varepsilon,$$

it follows that

$$\int_X \bar{f}_M\,\mathrm{d}\mu \le \int_X f\,\mathrm{d}\mu + 2\varepsilon + NM/L.$$

Since L may be chosen arbitrarily large and then ε arbitrarily small, we conclude that

$$\int_X \bar{f}_M\,\mathrm{d}\mu \le \int_X f\,\mathrm{d}\mu.$$

Now letting $M \to \infty$, we obtain

$$\int_X \bar{f}\,\mathrm{d}\mu \le \int_X f\,\mathrm{d}\mu.$$

The proof that

$$\int_X f\,\mathrm{d}\mu \le \int_X \underline{f}\,\mathrm{d}\mu$$

is similar. Given $\varepsilon > 0$, there exists for each $x \in X$ a positive integer n such that

$$n^{-1}\sum_{k=0}^{n-1} f(T^k x) \le \underline{f}(x) + \varepsilon. \qquad (**)$$

If F_n is the set of all $x \in X$ for which $(**)$ holds and if $E_n = \bigcup_{k=1}^{n} F_k$, we can choose N so large that

$$\int_{X\setminus E_N} f\,\mathrm{d}\mu < \varepsilon.$$

Put

$$\begin{aligned} \tilde{f}(x) &= f(x) \quad \text{if } x \in E_N, \\ &= 0 \quad \text{if } x \notin E_N. \end{aligned}$$

Let $\tau(x)$ be the least positive integer n for which $(**)$ holds if $x \in E_N$, and $\tau(x) = 1$ otherwise. The proof now goes through in the same way as before. □

It should be noticed that the preceding proof simplifies if the function f is bounded. In Birkhoff's original formulation the function f was the indicator function χ_B of an arbitrary set $B \in \mathscr{B}$. In this case the theorem says that, if $v_n(x)$ is the number of $k < n$ for which $T^k x \in B$, then $\lim_{n\to\infty} v_n(x)/n$ exists for almost all $x \in X$. That is, 'almost every point has an average sojourn time in any measurable set'.

A measure-preserving transformation T of the probability space $(X, \mathscr{B}, \mu)$ is said to be *ergodic* if, for every $B \in \mathscr{B}$ with $T^{-1}B = B$, either $\mu(B) = 0$ or $\mu(B) = 1$. Part (ii) of the next proposition says that this is the case if and only if 'time means and space means are equal'.

Proposition 18 *Let T be a measure-preserving transformation of the probability space $(X, \mathscr{B}, \mu)$. Then T is ergodic if and only if one of the following equivalent properties holds:*

(i) *if $f \in L(X, \mathscr{B}, \mu)$ satisfies $f(Tx) = f(x)$ almost everywhere, then f is constant almost everywhere;*

(ii) *if $f \in L(X, \mathscr{B}, \mu)$ then, for almost all $x \in X$,*

$$n^{-1}\sum_{k=0}^{n-1} f(T^k x) \to \int_X f \,\mathrm{d}\mu \quad \text{as } n \to \infty;$$

(iii) *if $A, B \in \mathscr{B}$, then*

$$n^{-1}\sum_{k=0}^{n-1} \mu(T^{-k}A \cap B) \to \mu(A)\mu(B) \quad \text{as } n \to \infty;$$

(iv) *if $C \in \mathscr{B}$ and $\mu(C) > 0$, then $\mu(\bigcup_{n\geq 1} T^{-n}C) = 1$;*

(v) *if $A, B \in \mathscr{B}$ and $\mu(A) > 0$, $\mu(B) > 0$, then $\mu(T^{-n}A \cap B) > 0$ for some $n > 0$.*

Proof Suppose first that T is ergodic and let $f \in L(X, \mathscr{B}, \mu)$ satisfy $f(Tx) = f(x)$ a.e. Put

$$\bar{f}(x) = \overline{\lim_{n\to\infty}}\, n^{-1}\sum_{k=0}^{n-1} f(T^k x).$$

Then $\bar{f}(Tx) = \bar{f}(x)$ for every $x \in X$ and $\bar{f}(x) = f(x)$ a.e. For any $\alpha \in \mathbb{R}$, let

$$A_\alpha = \{x \in X : \bar{f}(x) < \alpha\}.$$

Then $\mu(A_\alpha) = 0$ or 1, since $T^{-1}A_\alpha = A_\alpha$ and T is ergodic. Since $\mu(A_\alpha)$ is a nondecreasing function of α and $\mu(A_\alpha) \to 0$ as $\alpha \to -\infty$, $\mu(A_\alpha) \to 1$ as $\alpha \to +\infty$, there exists $\beta \in \mathbb{R}$ such that $\mu(A_\alpha) = 0$ for $\alpha < \beta$ and $\mu(A_\alpha) = 1$ for $\alpha > \beta$. It follows that $\mu(A_\beta) = 0$ and $\mu(B_\beta) = 1$, where

$$B_\beta = \{x \in X : \bar{f}(x) \leq \beta\}.$$

Hence $f(x) = \beta$ a.e. and (i) holds.

Suppose now that (i) holds and let $f \in L(X, \mathscr{B}, \mu)$. Then the function f^* in the statement of Theorem 17 must be constant a.e. Moreover, if γ is its constant value, we must have

$$\gamma = \int_X f^* \, \mathrm{d}\mu = \int_X f \, \mathrm{d}\mu.$$

Thus (i) implies (ii).

Suppose next that (ii) holds and let $A, B \in \mathscr{B}$. Then, for almost all $x \in X$,

$$\lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} \chi_A(T^k x) = \int_X \chi_A \, \mathrm{d}\mu = \mu(A).$$

Hence, for almost all $x \in X$,

$$\lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} \chi_A(T^k x)\chi_B(x) = \mu(A)\chi_B(x)$$

and so, by the dominated convergence theorem,

$$\begin{aligned}
\mu(A)\mu(B) &= \int_X \lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} \chi_A(T^k x)\chi_B(x) \, \mathrm{d}\mu(x) \\
&= \lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} \int_X \chi_A(T^k x)\chi_B(x) \, \mathrm{d}\mu(x) \\
&= \lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} \mu(T^{-k}A \cap B).
\end{aligned}$$

Thus (ii) implies (iii).

Suppose now that (iii) holds and choose $C \in \mathscr{B}$ with $\mu(C) > 0$. Put $A = \bigcup_{n\geq 0} T^{-n}C$ and $B = (\bigcup_{n\geq 1} T^{-n}C)^c$. Then, for every $k \geq 1$, $T^{-k}A \subseteq \bigcup_{n\geq 1} T^{-n}C$ and hence $\mu(T^{-k}A \cap B) = 0$. Thus

$$n^{-1} \sum_{k=0}^{n-1} \mu(T^{-k}A \cap B) = \mu(A \cap B)/n \to 0 \quad \text{as } n \to \infty.$$

Since $\mu(A) \geq \mu(C) > 0$, it follows from (iii) that $\mu(B) = 0$. Thus (iii) implies (iv).

Next choose any $A, B \in \mathscr{B}$ such that $\mu(A) > 0$, $\mu(B) > 0$. If (iv) holds, then $\mu(\bigcup_{n\geq 1} T^{-n}A) = 1$ and hence

$$\mu(B) = \mu\left(B \cap \bigcup_{n\geq 1} T^{-n}A\right) = \mu\left(\bigcup_{n\geq 1} (B \cap T^{-n}A)\right).$$

Since $\mu(B) > 0$, it follows that $\mu(B \cap T^{-n}A) > 0$ for some $n > 0$. Thus (iv) implies (v).

Finally choose $A \in \mathscr{B}$ with $T^{-1}A = A$ and put $B = A^c$. Then, for every $n \geq 1$, we have $\mu(T^{-n}A \cap B) = \mu(A \cap B) = 0$. If (v) holds, it follows that either $\mu(A) = 0$ or $\mu(B) = 0$. Hence (v) implies that T is ergodic. □

4 Applications

We now give some examples to illustrate the general concepts and results of the previous section.

(i) Suppose $X = \mathbb{R}^d/\mathbb{Z}^d$ is a d-dimensional torus, $\mathscr{B}$ is the family of *Borel subsets* of X (i.e., the σ-algebra of subsets generated by the family of open sets), and $\mu = \lambda$ is Lebesgue measure, i.e. $\mu(B) = \int_X \chi_B(x)\,\mathrm{d}x$ for any $B \in \mathscr{B}$, where χ_B is the indicator function of B. Every $x \in X$ is represented by a unique vector $(\xi_1, \ldots, \xi_d)$, where $0 \le \xi_k < 1$ $(k = 1, \ldots, d)$, and X is an abelian group with addition $z = x + y$ defined by $\zeta_k \equiv \xi_k + \eta_k \bmod 1$ $(k = 1, \ldots, d)$.

For any $a \in X$, the *translation* $T_a : X \to X$ defined by $T_a x = x + a$ is a measure-preserving transformation of the probability space $(X, \mathscr{B}, \lambda)$.

Proposition 19 *The translation $T_a : X \to X$ of the d-dimensional torus $X = \mathbb{R}^d/\mathbb{Z}^d$ is ergodic if and only if $1, \alpha_1, \ldots, \alpha_d$ are linearly independent over the rational field $\mathbb{Q}$, where $(\alpha_1, \ldots, \alpha_d)$ is the vector which represents a.*

Proof Suppose first that $1, \alpha_1, \ldots, \alpha_d$ are not linearly independent over $\mathbb{Q}$. Then there exists a nonzero vector $n \in \mathbb{Z}^d$ such that

$$n \cdot a = v_1\alpha_1 + \cdots + v_d\alpha_d \in \mathbb{Z}.$$

Hence if $f(x) = e(n \cdot x)$, then $f(T_a x) = f(x)$ for all x. Since f is not constant a.e., it follows from part (i) of Proposition 18 that T_a is not ergodic.

Suppose on the other hand that $1, \alpha_1, \ldots, \alpha_d$ are linearly independent over $\mathbb{Q}$ and let f be an integrable function such that $f(T_a x) = f(x)$ a.e. Then $f(T_a x)$ and $f(x)$ have the same Fourier coefficients:

$$\int_X f(x)e(-n \cdot x)\,\mathrm{d}x = \int_X f(x + a)e(-n \cdot x)\,\mathrm{d}x = e(n \cdot a)\int_X f(x)e(-n \cdot x)\,\mathrm{d}x.$$

Since $e(n \cdot a) \ne 1$ for all $n \ne 0$, it follows that

$$\int_X f(x)e(-n \cdot x)\,\mathrm{d}x = 0 \quad \text{for all } n \ne 0.$$

Since integrable functions with the same Fourier coefficients must agree almost everywhere, this proves that f is constant a.e. Hence, by Proposition 18 again, T_a is ergodic. □

If we compare Proposition 3′ and the remarks after its proof with Proposition 19, then we see from Theorems 1′-2′ and Proposition 18 that the following five statements are equivalent for $X = \mathbb{R}^d/\mathbb{Z}^d$ and any $a \in X$:

(α) the sequence $(\{na\})$ is dense in X;
(β) for every $x \in X$, the sequence $(x + na)$ is uniformly distributed in X;
(γ) the translation $T_a : X \to X$ is ergodic;
(δ) for each continuous function $f : X \to \mathbb{C}$, $\lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} f(T_a^k x) = \int_X f\,\mathrm{d}\lambda$ for all $x \in X$;

(ε) for each function $f \in L(X, \mathscr{B}, \lambda)$, $\lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} f(T_a^k x) = \int_X f \, d\lambda$ for almost all $x \in X$.

(ii) Again suppose $X = \mathbb{R}^d/\mathbb{Z}^d$ is a d-dimensional torus, $\mathscr{B}$ is the family of Borel subsets of X and $\mu = \lambda$ is Lebesgue measure. For any $d \times d$ matrix $A = (\alpha_{jk})$ of integers, let $R_A : X \to X$ be the map defined by $R_A x = x'$, where

$$\xi_j' \equiv \sum_{k=1}^{d} \alpha_{jk}\xi_k \mod 1 \ (j = 1, \ldots, d).$$

If $\det A = 0$ then R_A is not measure-preserving, since the image of $\mathbb{R}^d$ under A is contained in a hyperplane of $\mathbb{R}^d$. However, if $\det A \neq 0$ then R_A is measure-preserving, since each point of X is the image under R_A of $|\det A|$ distinct points of X, and a small region B of X is the image under R_A of $|\det A|$ disjoint regions, each with volume $|\det A|^{-1}$ times that of B. (This argument is certainly valid if A is a diagonal matrix, and the general case may be reduced to this by Proposition III.41.) Thus R_A is an *endomorphism* of the torus $\mathbb{R}^d/\mathbb{Z}^d$ if and only if A is nonsingular, and an *automorphism* if and only if $\det A = \pm 1$.

Proposition 20 *The endomorphism $R_A : X \to X$ of the d-dimensional torus $X = \mathbb{R}^d/\mathbb{Z}^d$ is ergodic if and only if no eigenvalue of the nonsingular matrix A is a root of unity.*

Proof For any $n \in \mathbb{Z}^d$ we have

$$e(n \cdot R_A x) = e(n \cdot Ax) = e(Dn \cdot x),$$

where $D = A^t$ is the transpose of A.

Suppose first that A, and hence also D, has an eigenvalue ω which is a root of unity: $\omega^p = 1$ for some positive integer p. Then $(D^p - I)z = 0$ for some nonzero vector z. Moreover, since D is a matrix of integers, we may assume that $z = m \in \mathbb{Z}^d$. We may further assume that $D^i m \neq D^j m$ for $0 \leq i < j < p$, by choosing p to have its least possible value. If we put

$$f(x) = e(m \cdot x) + e(m \cdot Ax) + \cdots + e(m \cdot A^{p-1}x),$$

then $f(R_A x) = f(x)$, but f is not constant a.e. Hence R_A is not ergodic, by Proposition 18.

Suppose next that R_A is not ergodic. Then, by Proposition 18 again, there exists a function $f \in L(X, \mathscr{B}, \lambda)$ such that $f(R_A x) = f(x)$ a.e., but $f(x)$ is not constant a.e. If the Fourier series of $f(x)$ is

$$\sum_{n\in\mathbb{Z}^d} c_n \, e(n \cdot x),$$

then the Fourier series of $f(R_A x)$ is

$$\sum_{n\in\mathbb{Z}^d} c_n \, e(n \cdot Ax) = \sum_{n\in\mathbb{Z}^d} c_n \, e(Dn \cdot x) = \sum_{n\in\mathbb{Z}^d} c_{D^{-1}n} \, e(n \cdot x)$$

and hence

$$c_n = c_{D^{-1}n} \quad \text{for every } n \in \mathbb{Z}^d.$$

But $c_m \neq 0$ for some nonzero $m \in \mathbb{Z}^d$, since f is not constant a.e., and $|c_n| \to 0$ as $|n| \to \infty$, since $f \in L(X, \mathscr{B}, \lambda)$. Since $c_{D^{-k}m} = c_m$ for every positive integer k, it follows that the subscripts $D^{-k}m$ are not all distinct. Hence $D^p m = m$ for some positive integer p and A has an eigenvalue which is a root of unity. □

(There are generalizations of Propositions 19 and 20 to translations and endomorphisms of any compact abelian group X, with Haar measure in place of Lebesgue measure.)

The preceding results have an application to the theory of 'normal numbers'. In fact, without any extra effort, we will consider also higher-dimensional generalizations. A vector $x \in \mathbb{R}^d$ is said to be *normal with respect to the matrix A*, where A is a $d \times d$ matrix of integers, if the sequence $(A^n x)$ is uniformly distributed mod 1.

Proposition 21 *Let A be a $d \times d$ matrix of integers. Then (λ-) almost all vectors $x \in \mathbb{R}^d$ are normal with respect to A if and only if A is nonsingular and no eigenvalue of A is a root of unity.*

Proof If A is nonsingular and no eigenvalue of A is a root of unity then, by Proposition 20, R_A is an ergodic measure-preserving transformation of the torus $X = \mathbb{R}^d/\mathbb{Z}^d$. Hence, by Proposition 18(ii), for each nonzero $m \in \mathbb{Z}^d$,

$$n^{-1} \sum_{k=0}^{n-1} e(m \cdot A^n x) \to 0 \quad \text{as } n \to \infty \quad \text{for almost all } x \in \mathbb{R}^d.$$

Since $\mathbb{Z}^d$ is countable, and the union of a countable number of sets of measure zero is again a set of measure zero, it follows that, for almost all $x \in \mathbb{R}^d$,

$$n^{-1} \sum_{k=0}^{n-1} e(m \cdot A^n x) \to 0 \quad \text{as } n \to \infty \quad \text{for every nonzero } m \in \mathbb{Z}^d.$$

Hence, by Theorem 2′, almost all $x \in \mathbb{R}^d$ are normal with respect to A.

If A is singular then, by the remark following the proof of Proposition 9, no $x \in \mathbb{R}^d$ is normal with respect to A. Suppose finally that some eigenvalue of A is a root of unity. Then there exists a positive integer p and a nonzero vector $z \in \mathbb{Z}^d$ such that $D^p z = z$, where $D = A^t$. If

$$f(x) = e(z \cdot x) + e(z \cdot Ax) + \cdots + e(z \cdot A^{p-1}x),$$

then $f(Ax) = f(x)$ and hence

$$n^{-1} \sum_{k=0}^{n-1} f(A^k x) = f(x).$$

But if x is normal with respect to A then, by Theorem 1′,

$$n^{-1}\sum_{k=0}^{n-1} f(A^k x) \to \int_X f \, d\lambda = 0.$$

Since f is not zero a.e., it follows that the set of all x which are normal with respect to A does not have full measure. □

We consider next when normality with respect to one matrix coincides with normality with respect to another matrix.

Proposition 22 *Let A be a $d \times d$ nonsingular matrix of integers, no eigenvalue of which is a root of unity. Then, for any positive integer q, the vector $x \in \mathbb{R}^d$ is normal with respect to A^q if and only if it is normal with respect to A.*

Proof It follows at once from Proposition 9 that if x is normal with respect to A, then it is also normal with respect to A^q.

Suppose, on the other hand, that x is normal with respect to A^q. Then, by Theorem 2′, for every nonzero vector $m \in \mathbb{Z}^d$,

$$N^{-1}\sum_{n=0}^{N-1} e(m \cdot A^{nq} x) \to 0 \quad \text{as } N \to \infty.$$

Put $D = A^t$. Since D is a nonsingular matrix of integers, $D^j m$ is a nonzero vector in $\mathbb{Z}^d$ for any integer $j \geq 0$ and hence

$$N^{-1}\sum_{n=0}^{N-1} e(m \cdot A^{nq+j} x) = N^{-1}\sum_{n=0}^{N-1} e(D^j m \cdot A^{nq} x) \to 0 \quad \text{as } N \to \infty.$$

Adding these relations for $j = 0, 1, \ldots, q-1$ and dividing by q, we obtain

$$(Nq)^{-1}\sum_{n=0}^{Nq-1} e(m \cdot A^n x) \to 0 \quad \text{as } N \to \infty.$$

Since the sum of at most q terms $e(m \cdot A^n x)$ has absolute value at most q it follows that, also without restricting N to be a multiple of q,

$$N^{-1}\sum_{n=0}^{N-1} e(m \cdot A^n x) \to 0 \quad \text{as } N \to \infty.$$

Hence, by Theorem 2′, x is normal with respect to A. □

Corollary 23 *Let A be a $d \times d$ nonsingular integer matrix, no eigenvalue of which is a root of unity, and let B be a $d \times d$ integer matrix such that $A^p = B^q$ for some positive integers p, q. Then $x \in \mathbb{R}^d$ is normal with respect to A if and only if x is normal with respect to B.*

Proof This follows at once from Proposition 22, since the hypotheses imply that also B is nonsingular and has no eigenvalue which is a root of unity. □

Brown and Moran (1993) have shown, conversely, that if A, B are *commuting* $d \times d$ nonsingular integer matrices, no eigenvalues of which are roots of unity, such that the set of all vectors normal with respect to A coincides with the set of all vectors normal with respect to B, then $A^p = B^q$ for some positive integers p, q.

These results will now be specialized to the scalar case. A real number x is said to be *normal to the base* a, where a is an integer ≥ 2, if the sequence $(a^n x)$ is uniformly distributed mod 1. It is readily shown that x is normal to the base a if and only if, in the expansion of x to the base a:

$$x = \lfloor x \rfloor + x_1/a + x_2/a^2 + \cdots,$$

where $x_i \in \{0, 1, \ldots, a-1\}$ for all $i \geq 1$ and $x_i = a - 1$ for at most finitely many i, every block of digits occurs with the proper frequency; i.e., for any positive integer k and any $a_1, \ldots, a_k \in \{0, 1, \ldots, a-1\}$, the number $v(N)$ of i with $1 \leq i \leq N$ such that

$$x_i = a_1, x_{i+1} = a_2, \ldots, x_{i+k-1} = a_k,$$

satisfies $v(N)/N \to a^{-k}$ as $N \to \infty$. By Proposition 21, almost all real numbers x are normal to a given base a. The original proof of this by Borel (1909) was a forerunner of Birkhoff's ergodic theorem. (In fact Borel's proof was faulty, but his paper was influential. Borel used a different definition of normal number, but Wall (1949) showed that it was equivalent to the definition in terms of uniform distribution adopted here.)

The first published proof of the scalar case of Corollary 23 was given by Schmidt (1960), who also proved the scalar version of the result of Brown and Moran: the set of all numbers normal to the base a coincides with the set of all numbers normal to the base b, where a and b are integers ≥ 2, if and only if $a^p = b^q$ for some positive integers p, q.

Although almost all real numbers are normal to *every* base a, it is still not known if such familiar irrational numbers as $\sqrt{2}, e$ or π are normal to some base. There are, however, various explicit constructions of normal numbers. In particular, Champernowne (1933) showed that the real number θ whose expansion to the base 10 is composed of the positive integers in their natural order, in other words, $\theta = 0.123456789101112\ldots$, is itself normal to the base 10.

(iii) Let A be a set of finite cardinality r, which for definiteness we take to be the set $\{1, \ldots, r\}$, and let $p_1, \ldots, p_r$ be positive real numbers with sum 1. If $\mathscr{B}_0$ is the family of all subsets of the finite set A and if, for any $B_0 \in \mathscr{B}_0$, we put $\mu_0(B_0) = \sum_{a \in B_0} p_a$, then μ_0 is a probability measure and $(A, \mathscr{B}_0, \mu_0)$ is a probability space.

Now let X be the set of all bi-infinite sequences $x = (\ldots, x_{-2}, x_{-1}, x_0, x_1, x_2, \ldots)$ with $x_i \in A$ for every $i \in \mathbb{Z}$. Thus X is the product of infinitely many copies of A. We construct a *product measure* on X in the following way.

For any finite sequence $(a_{-m}, \ldots, a_0, \ldots, a_m)$ with $a_i \in A$ for $-m \leq i \leq m$, define the (special) *cylinder set* $[a_{-m}, \ldots, a_m]$ of *order* m to be the set of all $x \in X$ such that $x_i = a_i$ for $-m \leq i \leq m$. There are r^{2m+1} distinct cylinder sets of order m, distinct cylinder sets are disjoint and X is the union of them all.

Let $\mathscr{C}_m$ denote the collection of all unions of distinct cylinder sets of order m. Thus $X \in \mathscr{C}_m$ and, if $B \in \mathscr{C}_m$, then $B^c = X \backslash B \in \mathscr{C}_m$. Moreover $B, C \in \mathscr{C}_m$ implies $B \cup C \in \mathscr{C}_m$ and $B \cap C \in \mathscr{C}_m$. If $B \in \mathscr{C}_m$, say

$$B = [a_{-m}, \ldots, a_m] \cup \cdots \cup [a'_{-m}, \ldots, a'_m],$$

we define

$$\mu_m(B) = p_{a_{-m}} \cdots p_{a_m} + \cdots + p_{a'_{-m}} \cdots p_{a'_m}.$$

Then $\mu_m(X) = 1$, $\mu_m(B) \geq 0$ for every $B \in \mathscr{C}_m$, and

$$\mu_m(B \cup C) = \mu_m(B) + \mu_m(C) \text{ if } B, C \in \mathscr{C}_m \text{ and } B \cap C = \emptyset.$$

Every union of cylinder sets of order m is also a union of cylinder sets of order $m+1$, since

$$[a_{-m}, \ldots, a_m] = \bigcup_{a,a' \in A} [a, a_{-m}, \ldots, a_m, a'].$$

Thus $\mathscr{C}_m \subseteq \mathscr{C}_{m+1}$. Moreover μ_{m+1} continues μ_m, since

$$\begin{aligned}\mu_{m+1}([a_{-m}, \ldots, a_m]) &= \sum_{j,j'=1}^{r} p_j p_{j'} p_{a_{-m}} \cdots p_{a_m} \\ &= \mu_m([a_{-m}, \ldots, a_m]) \Big(\sum_{j=1}^{r} p_j\Big)\Big(\sum_{j'=1}^{r} p_{j'}\Big) \\ &= \mu_m([a_{-m}, \ldots, a_m]).\end{aligned}$$

Let μ denote the continuation of all μ_m to $\mathscr{C} = \mathscr{C}_0 \cup \mathscr{C}_1 \cup \ldots$. If $B, C \in \mathscr{C}$, then $B, C \in \mathscr{C}_m$ for some m. Hence, for given $C \in \mathscr{C}$, there are only finitely many distinct $B \in \mathscr{C}$ such that $B \subseteq C$. Consequently, if C is the union of a sequence of disjoint sets $C_n \in \mathscr{C} (n = 1, 2, \ldots)$, then $C_n = \emptyset$ for all large n and $\mu(C) = \sum_{n \geq 1} \mu(C_n)$. It follows, by a construction due to Carathéodory (1914), that μ can be uniquely extended to the σ-algebra $\mathscr{B}$ of subsets of X generated by $\mathscr{C}$ so that $(X, \mathscr{B}, \mu)$ is a probability space. For any $\varepsilon > 0$ there exists, for each $B \in \mathscr{B}$, some $C \in \mathscr{C}$ such that $\mu(B \Delta C) < \varepsilon$.

The *two-sided Bernoulli shift* $B_{p_1,\ldots,p_r}$ is the map $\sigma : X \to X$ defined by $\sigma x = x'$, where $x'_i = x_{i+1}$ for every $i \in \mathbb{Z}$. It is a measure-preserving transformation of the probability space $(X, \mathscr{B}, \mu)$, since

$$\sigma^{-1}[a_{-m}, \ldots, a_m] = \bigcup_{a,a' \in A} [a, a', a_{-m}, \ldots, a_m]$$

and hence

$$\begin{aligned}\mu(\sigma^{-1}[a_{-m}, \ldots, a_m]) &= \sum_{j,j'=1}^{r} p_j p_{j'} p_{a_{-m}} \cdots p_{a_m} \\ &= \sum_{j,j'=1}^{r} p_j p_{j'} \mu([a_{-m}, \ldots, a_m]) = \mu([a_{-m}, \ldots, a_m]).\end{aligned}$$

The Bernoulli shift $B_{1/2,1/2}$ is a model for the random process consisting of bi-infinite sequences of coin-tossings.

We may define the *general cylinder set* $C_{i_1\dots i_k}^{a_1\dots a_k}$, where $i_1, \dots, i_k$ are distinct integers, to be the set of all $x \in X$ such that

$$x_{i_1} = a_1, \dots, x_{i_k} = a_k.$$

In particular, $C_i^a = \sigma^{-i}[a]$ and hence $\mu(C_i^a) = p_a$. It follows by induction on k that

$$\mu(C_{i_1\dots i_k}^{a_1\dots a_k}) = p_{a_1} \cdots p_{a_k}.$$

Proposition 24 *For any given positive numbers $p_1, \dots, p_r$ with sum 1, the two-sided Bernoulli shift $B_{p_1,\dots,p_r}$ is ergodic.*

Proof Suppose $B \in \mathscr{B}$ and $\sigma^{-1}B = B$. For any $\varepsilon > 0$ there exists a set $C \in \mathscr{C}$ such that

$$\mu(B \,\Delta C) = \mu(B\backslash C) + \mu(C\backslash B) < \varepsilon.$$

Then

$$\begin{aligned} |\mu(B) - \mu(C)| &= |\mu(C \cap B) + \mu(B\backslash C) - \mu(C \cap B) - \mu(C\backslash B)| \\ &\le \mu(B\backslash C) + \mu(C\backslash B) < \varepsilon \end{aligned}$$

and hence

$$|\mu(B)^2 - \mu(C)^2| = \{\mu(C) + \mu(B)\}\,|\mu(B) - \mu(C)| < 2\varepsilon.$$

We may suppose that C is the union of finitely many special cylinder sets of order m. Since

$$\sigma^{-n}[a_{-m}, \dots, a_m] = C_{-m+n,\dots,m+n}^{a_{-m},\dots,a_m},$$

for $n > 2m$ we have

$$[a'_{-m}, \dots, a'_m] \cap \sigma^{-n}[a_{-m}, \dots, a_m] = C_{-m,\dots,m,-m+n,\dots,m+n}^{a'_{-m},\dots,a'_m,a_{-m},\dots,a_m},$$

and hence

$$\begin{aligned} \mu([a'_{-m}, \dots, a'_m] \cap \sigma^{-n}[a_{-m}, \dots, a_m]) &= p_{a'_{-m}} \cdots p_{a'_m} p_{a_{-m}} \cdots p_{a_m}, \\ &= \mu([a'_{-m}, \dots, a'_m])\mu([a_{-m}, \dots, a_m]). \end{aligned}$$

It follows that if $n > 2m$, then

$$\mu(C \cap \sigma^{-n}C) = \mu(C)^2.$$

But

$$\mu(B\backslash(C \cap \sigma^{-n}C)) \le 2\mu(B\backslash C),$$

since

$$B\backslash(C \cap \sigma^{-n}C) \subseteq (B\backslash C) \cup (B\backslash\sigma^{-n}C) \subseteq (B\backslash C) \cup \sigma^{-n}(B\backslash C),$$

and similarly

$$\mu((C \cap \sigma^{-n}C)\backslash B) \leq 2\mu(C\backslash B).$$

Hence

$$|\mu(B) - \mu(C \cap \sigma^{-n}C)| \leq \mu(B\backslash(C \cap \sigma^{-n}C)) + \mu((C \cap \sigma^{-n}C)\backslash B) < 2\varepsilon.$$

Thus

$$\begin{aligned} 0 \leq \mu(B) - \mu(B)^2 &= \mu(B) - \mu(C \cap \sigma^{-n}C) + \mu(C \cap \sigma^{-n}C) - \mu(B)^2 \\ &< 2\varepsilon + \mu(C)^2 - \mu(B)^2 < 4\varepsilon. \end{aligned}$$

Since ε is arbitrary, we conclude that $\mu(B) = \mu(B)^2$. Hence $\mu(B) = 0$ or 1, and σ is ergodic. □

Similarly, if Y is the set of all infinite sequences $y = (y_1, y_2, y_3, \ldots)$ with $y_i \in A$ for every $i \in \mathbb{N}$, then the *one-sided Bernoulli shift* $B^+_{p_1,\ldots,p_r}$, i.e. the map $\tau : Y \to Y$ defined by $\tau y = y'$, where $y'_i = y_{i+1}$ for every $i \in \mathbb{N}$, is a measure-preserving transformation of the analogously constructed probability space $(Y, \mathscr{B}, \mu)$. It should be noted that, although $\tau Y = Y$, τ is not invertible. In the same way as for the two-sided shift, it may be shown that the one-sided Bernoulli shift $B^+_{p_1,\ldots,p_r}$ is always ergodic.

(iv) An example of some historical interest is the 'continued fraction' or *Gauss* map. Let $X = [0, 1]$ be the unit interval and $T : X \to X$ the map defined (in the notation of §1) by

$$\begin{aligned} T\xi &= \{\xi^{-1}\} \quad \text{if } \xi \in (0, 1), \\ &= 0 \qquad \text{if } \xi = 0 \text{ or } 1. \end{aligned}$$

Thus T acts as the shift operator on the continued fraction expansion of ξ: if

$$\xi = [0; a_1, a_2, \ldots] = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cdots}},$$

then $T\xi = [0; a_2, a_3, \ldots]$. (In the terminology of Chapter IV, the complete quotients of ξ are $\xi_{n+1} = 1/T^n\xi$.)

It is not difficult to show that T is a measure-preserving transformation of the probability space $(X, \mathscr{B}, \mu)$, where $\mathscr{B}$ is the family of Borel subsets of $X = [0, 1]$ and μ is the 'Gauss' measure defined by

$$\mu(B) = (\log 2)^{-1} \int_B (1 + x)^{-1}\, dx.$$

It may further be shown that T is ergodic. Hence, by Birkhoff's ergodic theorem, if f is an integrable function on the interval X then, for almost all $\xi \in X$,

$$\lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} f(T^k\xi) = (\log 2)^{-1} \int_X f(x)(1 + x)^{-1}\, dx.$$

Here it makes no difference if 'integrable' and 'almost all' refer to the invariant measure μ or to Lebesgue measure, since $1/2 \le (1+x)^{-1} \le 1$.

Taking f to be the indicator function of the set $\{\xi \in X : a_1 = m\}$, we see that the asymptotic relative frequency of the positive integer m among the partial quotients $a_1, a_2, \ldots$ is almost always

$$(\log 2)^{-1} \int_{(m+1)^{-1}}^{m^{-1}} (1+x)^{-1}\, dx = (\log 2)^{-1} \log((m+1)^2/(m(m+2))).$$

It follows, in particular, that almost all $\xi \in X$ have unbounded partial quotients.

Again, by taking $f(\xi) = \log \xi$ it may be shown that, for almost all $\xi \in X$,

$$\lim_{n\to\infty} (1/n) \log q_n(\xi) = \pi^2/(12 \log 2),$$

where $q_n(\xi)$ is the denominator of the n-th convergent p_n/q_n of ξ. This was first proved by Lévy (1929).

In a letter to Laplace, Gauss (1812) stated that, for each $x \in (0, 1)$, the proportion of $\xi \in X$ for which $T^n \xi < x$ converges as $n \to \infty$ to $\log(1+x)/(\log 2)$ and he asked if Laplace could provide an estimate for the rapidity of convergence. If one writes

$$r_n(x) = m_n(x) - \log(1+x)/(\log 2),$$

where $m_n(x)$ is the Lebesgue measure of the set of all $\xi \in X$ such that $T^n \xi < x$, then Gauss's statement is that $r_n(x) \to 0$ as $n \to \infty$ and his question is, how fast?

Gauss's statement was first proved by Kuz'min (1928), who also gave an estimate for the rapidity of convergence. If one regards Gauss's statement as a proposition in ergodic theory, then one needs to know that T is not only ergodic but even *mixing*, i.e. for all $A, B \in \mathscr{B}$,

$$\mu(T^{-n} A \cap B) \to \mu(A)\mu(B) \quad \text{as } n \to \infty.$$

Kuz'min's estimate $r_n(x) = O(q^{\sqrt{n}})$ for some $q \in (0, 1)$ was improved by Lévy (1929) and Szüsz (1961) to $r_n(x) = O(q^n)$ with $q = 0.7$ and $q = 0.485$ respectively. A substantial advance was made by Wirsing (1974). By means of an infinite-dimensional generalization of a theorem of Perron (1907) and Frobenius (1908) on positive matrices, he showed that

$$r_n(x) = (-\lambda)^n \psi(x) + O(x(1-x)\mu^n),$$

where ψ is a twice continuously differentiable function with $\psi(0) = \psi(1) = 0$, $0 < \mu < \lambda$ and $\lambda = 0.303663\ldots$. Wirsing's analysis has been extended by Babenko (1978) and Mayer (1990).

(v) Suppose we are given a system of ordinary differential equations

$$dx/dt = f(x), \qquad (\dagger)$$

where $x \in \mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}^d$ is a continuously differentiable function. Then, for any $x \in \mathbb{R}^d$, there is a unique solution $\varphi_t(x)$ of $(\dagger)$ such that $\varphi_0(x) = x$.

Suppose further that there exists an *invariant region* $X \subseteq \mathbb{R}^d$. That is, X is the closure of a bounded connected open set and $x \in X$ implies $\varphi_t(x) \in X$. Then the map $T_t\colon X \to X$ given by $T_t x = \varphi_t(x)$ is defined for every $t \in \mathbb{R}$ and satisfies $T_{t+s}x = T_t(T_s x)$.

Suppose finally that $\operatorname{div} f = 0$ for every $x \in \mathbb{R}^d$, where $x = (x_1, \ldots, x_d)$, $f = (f_1, \ldots, f_d)$ and

$$\operatorname{div} f := \sum_{k=1}^{d} \partial f_k / \partial x_k.$$

Then, by a theorem due to Liouville, the map T_t sends an arbitrary region into a region of the same volume. (For the statement and proof of Liouville's theorem see, for example, V.I. Arnold, *Mathematical methods of classical mechanics*, Springer-Verlag, New York, 1978.) It follows that if $\mathscr{B}$ is the family of Borel subsets of X and μ Lebesgue measure, normalized so that $\mu(X) = 1$, then T_t is a measure-preserving transformation of the probability space $(X, \mathscr{B}, \mu)$.

An important special case is the Hamiltonian system of ordinary differential equations

$$dp_i/dt = -\partial H/\partial q_i, \quad dq_i/dt = \partial H/\partial p_i \quad (i = 1, \ldots, n),$$

where $H(p_1, \ldots, p_n, q_1, \ldots, q_n)$ is a twice continuously differentiable real-valued function. The divergence does indeed vanish identically in this case, since

$$-\sum_{i=1}^{n} \partial^2 H/\partial p_i \partial q_i + \sum_{i=1}^{n} \partial^2 H/\partial q_i \partial p_i = 0.$$

Furthermore, for any $h \in \mathbb{R}$, the energy surface $X\colon H(p, q) = h$ is invariant, since

$$dH[p(t), q(t)]/dt = \sum_{i=1}^{n} \partial H/\partial p_i(-\partial H/\partial q_i) + \sum_{i=1}^{n} \partial H/\partial q_i \partial H/\partial p_i = 0.$$

It is not difficult to show that if σ is the volume element on X induced by the Euclidean metric $\| \ \|$ on $\mathbb{R}^{2n}$, and if

$$\nabla H = (\partial H/\partial p_1, \ldots, \partial H/\partial p_n, \partial H/\partial q_1, \ldots, \partial H/\partial q_n)$$

is the gradient of H, then the maps T_t preserve the measure μ on X defined by

$$\mu(B) = \int_B \mathrm{d}\sigma / \|\nabla H\|.$$

If X is compact, this measure can be normalized and we obtain a family of measure-preserving transformations $T_t (t \in \mathbb{R})$ of the corresponding probability space.

(vi) Many problems arising in mechanics may be reduced by a change of variables to the geometric problem of *geodesic flow*. If M is a smooth Riemannian manifold then the set of all pairs (x, v), where $x \in M$ and v is a unit vector in the tangent space to

M at x, can be given the structure of a Riemannian manifold, the *unit tangent bundle* T_1M. Evidently T_1M is a $(2n-1)$-dimensional manifold if M is n-dimensional. There is a natural measure μ on T_1M such that $\mathrm{d}\mu = \mathrm{d}v_q\,\mathrm{d}\omega_q$, where $\mathrm{d}v_q$ is the volume element at q of the Riemannian manifold M and ω_q is Lebesgue measure on the unit sphere S^{n-1} in the tangent space to M at x. If M is compact, then the measure μ can be normalized so that $\mu(T_1M) = 1$.

A *geodesic* on M is a curve $\gamma \subseteq M$ such that the length of every curve in M joining a point $x \in \gamma$ to any sufficiently close point $y \in \gamma$ is not less than the length of the arc of γ which joins x and y. Given any point $(x, v) \in T_1M$, there is a unique geodesic passing through x in the direction of v. The geodesic flow on T_1M is the flow $\varphi_t\colon T_1M \to T_1M$ defined by $\varphi_t(x, v) = (x_t, v_t)$, where x_t is the point of M reached from x after time t by travelling with unit speed along the geodesic determined by (x, v) and v_t is the unit tangent vector to this geodesic at x_t. If M is compact then, for every real t, φ_t is defined and is a measure-preserving transformation of the corresponding probability space $(T_1M, \mathscr{B}, \mu)$.

The geodesics on a compact 2-dimensional manifold M whose curvature at each point is negative were profoundly studied by Hadamard (1898). It was first shown by E. Hopf (1939) that in this case φ_t is ergodic for every $t > 0$. (We must exclude $t = 0$, since φ_0 is the identity map.) This result has been considerably generalized by Anosov (1967) and others. In particular, the geodesic flow on a compact n-dimensional Riemannian manifold is ergodic if at each point the curvature of every 2-dimensional section is negative.

Although the preceding examples look quite different, some of them are not 'really' different, i.e. apart from sets of measure zero. More precisely, if $(X_1, \mathscr{B}_1, \mu_1)$ and $(X_2, \mathscr{B}_2, \mu_2)$ are probability spaces with measure-preserving transformations $T_1\colon X_1 \to X_1$ and $T_2\colon X_2 \to X_2$, we say that T_1 is *isomorphic* to T_2 if there exist sets $X_1' \in \mathscr{B}_1$, $X_2' \in \mathscr{B}_2$ with $\mu_1(X_1') = 1$, $\mu_2(X_2') = 1$ and $T_1X_1' \subseteq X_1'$, $T_2X_2' \subseteq X_2'$, and a bijective map φ of X_1' onto X_2' such that

(i) for any $B_1 \subseteq X_1'$, $B_1 \in \mathscr{B}_1$ if and only if $\varphi(B_1) \in \mathscr{B}_2$ and then $\mu_1(B_1) = \mu_2(\varphi(B_1))$;
(ii) $\varphi(T_1x) = T_2\varphi(x)$ for every $x \in X_1'$.

For example, it is easily shown that the Bernoulli shift $B_{p_1,\dots,p_r}$ is isomorphic to the following transformation of the unit square, equipped with Lebesgue measure. Divide the square into r vertical strips of width $p_1, \dots, p_r$; then contract the height of the i-th strip and expand its width so that it has height p_i and width 1; finally combine these rectangles to form the unit square again by regarding them as horizontal strips of height $p_1, \dots, p_r$. (For $r = 2$ and $p_1 = p_2 = 1/2$, this transformation of the unit square is allegedly used by bakers when kneading dough.)

It is easily shown also that isomorphism is an equivalence relation and that it preserves ergodicity. However, it is usually quite difficult to show that two measure-preserving transformations are indeed isomorphic. A period of rapid growth was initiated with the definition by Kolmogorov (1958), and its practical implementation by Sinai (1959), of a new numerical isomorphism invariant, the *entropy* of a measure-preserving transformation. For the formal definition of entropy we refer to the texts on ergodic theory cited at the end of the chapter. Here we merely state its value for some of the preceding examples.

Any translation T_a of the torus $\mathbb{R}^d/\mathbb{Z}^d$ has entropy zero, whereas the endomorphism R_A of $\mathbb{R}^d/\mathbb{Z}^d$ has entropy

$$\sum_{i:\,|\lambda_i|>1} \log|\lambda_i|,$$

where $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of the matrix A and the summation is over those of them which lie outside the unit circle.

The two-sided Bernoulli shift $B_{p_1,\ldots,p_r}$ has entropy

$$-\sum_{j=1}^{r} p_j \log p_j,$$

and the entropy of the one-sided Bernoulli shift $B^+_{p_1,\ldots,p_r}$ is given by the same formula. It follows that $B_{1/2,1/2}$ is not isomorphic to $B_{1/3,1/3,1/3}$, since the first has entropy $\log 2$ and the second has entropy $\log 3$. Ornstein (1970) established the remarkable result that two-sided Bernoulli shifts are completely classified by their entropy: $B_{p_1,\ldots,p_r}$ is isomorphic to $B_{q_1,\ldots,q_s}$ if and only if

$$-\sum_{j=1}^{r} p_j \log p_j = -\sum_{k=1}^{s} q_k \log q_k.$$

This is no longer true for one-sided Bernoulli shifts. Walters (1973) has shown that $B^+_{p_1,\ldots,p_r}$ is isomorphic to $B^+_{q_1,\ldots,q_s}$ if and only if $r = s$ and $q_1, \ldots, q_s$ is a permutation of $p_1, \ldots, p_r$.

The Gauss map $Tx = \{x^{-1}\}$ has entropy $\pi^2/6 \log 2$. Although it is mixing, it is not isomorphic to a Bernoulli shift.

Katznelson (1971) showed that any ergodic automorphism of the torus $\mathbb{R}^d/\mathbb{Z}^d$ is isomorphic to a two-sided Bernoulli shift, and Lind (1977) has extended this result to ergodic automorphisms of any compact abelian group.

Ornstein and Weiss (1973) showed that, if φ_t is the geodesic flow on a smooth (of class C^3) compact two-dimensional Riemannian manifold whose curvature at each point is negative, then φ_t is isomorphic to a two-sided Bernoulli shift for every $t > 0$. Although, as Hilbert showed, a compact surface of negative curvature cannot be imbedded in $\mathbb{R}^3$, the geodesic flow on a surface of negative curvature can be realized as the motion of a particle constrained to move on a surface in $\mathbb{R}^3$ subject to centres of attraction and repulsion in the ambient space. The isomorphism with a Bernoulli shift shows that a deterministic mechanical system can generate a random process. Thus philosophical objections to 'Laplacian determinism' or to 'God playing dice' do not seem to have much point.

5 Recurrence

It was shown by Poincaré (1890) that the paths of a Hamiltonian system of differential equations almost always return to any neighbourhood, however small, of their initial

points. Poincaré's proof was inevitably incomplete, since at the time measure theory did not exist. However, Carathéodory (1919) showed that his argument could be made rigorous with the aid of Lebesgue measure:

Proposition 25 *Let $T: X \to X$ be a measure-preserving transformation of the probability space $(X, \mathscr{B}, \mu)$. Then almost all points of any $B \in \mathscr{B}$ return to B infinitely often, i.e. for each $x \in B$, apart from a set of μ-measure zero, there exists an increasing sequence (n_k) of positive integers such that $T^{n_k}x \in B$ $(k = 1, 2, \ldots)$.*

Furthermore, if $\mu(B) > 0$, then $\mu(B \cap T^{-n}B) > 0$ for infinitely many $n \geq 1$.

Proof For any $N \geq 0$, put $B_N = \bigcup_{n \geq N} T^{-n}B$. Then

$$A := \bigcap_{N \geq 0} B_N$$

is the set of all points $x \in X$ such that $T^n x \in B$ for infinitely many positive integers n. Since $B_{N+1} = T^{-1}B_N$, we have $\mu(B_{N+1}) = \mu(B_N)$ and hence $\mu(B_N) = \mu(B_0)$ for all $N \geq 1$. Since $B_{N+1} \subseteq B_N$, it follows that

$$\mu(A) = \lim_{N \to \infty} \mu(B_N) = \mu(B_0).$$

Since $A \subseteq B_0$, this implies

$$\mu(B_0 \backslash A) = \mu(B_0) - \mu(A) = 0$$

and hence, since $B \subseteq B_0$, $\mu(B \backslash A) = 0$.

This proves the first statement of the proposition. If $\mu(B \cap T^{-n}B) = 0$ for all $n \geq m$, then $\mu(B \cap B_N) = 0$ for all $N \geq m$ and hence

$$\mu(B \cap A) = \lim_{N \to \infty} \mu(B \cap B_N) = 0.$$

Consequently

$$\mu(B) = \mu(B \backslash A) + \mu(B \cap A) = 0,$$

which proves the second statement of the proposition. □

Furstenberg (1977) extended Proposition 25 in the following way:

Let T be a measure-preserving transformation of the probability space $(X, \mathscr{B}, \mu)$. If $B \in \mathscr{B}$ with $\mu(B) > 0$ and if $p \geq 2$, then $\mu(B \cap T^{-n}B \cap \cdots \cap T^{-(p-1)n}B) > 0$ for some $n \geq 1$.

His proof of this theorem made heavy use of ergodic theory and, in particular, of a new structure theory for measure-preserving transformations. From his theorem he was able to deduce quite easily a result for which Szemeredi (1975) had given a complicated combinatorial proof:

Let S be a subset of the set $\mathbb{N}$ of positive integers which has positive upper density; i.e., for some $\alpha \in (0, 1)$, there exist arbitrarily long intervals $I \subseteq \mathbb{N}$ containing at least $\alpha|I|$ elements of S. Then S contains arithmetic progressions of arbitrary finite length.

Furstenberg's approach to this result is not really shorter than Szemeredi's, but it is much more systematic. In fact the following generalization of Furstenberg's theorem was given soon afterwards by Furstenberg and Katznelson (1978):

If $T_1, \ldots, T_p$ are commuting measure-preserving transformations of the probability space $(X, \mathscr{B}, \mu)$ and if $B \in \mathscr{B}$ with $\mu(B) > 0$, then $\mu(B \cap T_1^{-n} B \cap \cdots \cap T_p^{-n} B) > 0$ for infinitely many $n \geq 1$.

Furstenberg and Katznelson could then deduce quite easily a multi-dimensional extension of Szemeredi's theorem which is still beyond the reach of combinatorial methods. Szemeredi's theorem was itself a far-reaching generalization of a famous theorem of van der Waerden (1927):

If $\mathbb{N} = S_1 \cup \cdots \cup S_r$ is a partition of the set of all positive integers into finitely many subsets, then one of the subsets S_j contains arithmetic progressions of arbitrary finite length.

Szemeredi's result further indicates how the subset S_j should be chosen.

Poincaré's measure-theoretic recurrence theorem has a topological counterpart due to Birkhoff (1912):

If X is a compact metric space and $T : X \to X$ a continuous map, then there exists a point $z \in X$ and an increasing sequence (n_k) of positive integers such that $T^{n_k} z \to z$ as $k \to \infty$.

Before Furstenberg and Katznelson proved their measure-theoretic theorem, Furstenberg and Weiss (1978) had already proved its topological counterpart:

If X is a compact metric space and $T_1, \ldots, T_p$ commuting continuous maps of X into itself, then there exists a point $z \in X$ and an increasing sequence (n_k) of positive integers such that $T_i^{n_k} z \to z$ as $k \to \infty$ $(i = 1, \ldots, p)$.

From their theorem Furstenberg and Weiss were able to deduce quite easily both van der Waerden's theorem and a known multi-dimensional generalization of it, due to Grünwald. It would take too long to prove here Szemeredi's theorem by the method of Furstenberg and Katznelson, but we will prove van der Waerden's theorem by the method of Furstenberg and Weiss. The proof illustrates how results in one area of mathematics can find application in another area which is apparently unrelated.

Proposition 26 *Let (X,d) be a compact metric space and $T : X \to X$ a continuous map. Then, for any real $\varepsilon > 0$ and any $p \in \mathbb{N}$, there exists some $z \in X$ and $n \in \mathbb{N}$ such that*

$$\mathrm{d}(T^n z, z) < \varepsilon, \quad \mathrm{d}(T^{2n} z, z) < \varepsilon, \ldots, \mathrm{d}(T^{pn} z, z) < \varepsilon.$$

Proof (i) A subset A of X is said to be *invariant* under T if $TA \subseteq A$. The closure $\bar{A}$ of an invariant set A is again invariant since, by the continuity of T, $T\bar{A} \subseteq \overline{TA}$. Let $\mathscr{F}$ be the collection of all nonempty closed invariant subsets of X. Clearly $\mathscr{F}$ is not empty, since $X \in \mathscr{F}$. If we regard $\mathscr{F}$ as partially ordered by inclusion then, by *Hausdorff's maximality theorem*, $\mathscr{F}$ contains a maximal totally ordered subcollection $\mathscr{T}$. The intersection Z of all the subsets in $\mathscr{T}$ is both closed and invariant. It is also nonempty, since X is compact. Hence $Z \in \mathscr{T}$ and, by construction, no nonempty proper closed subset of Z is invariant.

By replacing X by its compact subset Z we may now assume that the only closed invariant subsets of X itself are X and $\emptyset$.

(ii) For any given $z \in X$, the closure of the set $(T^n z)_{n\geq 1}$ is a nonempty closed invariant subset of X and therefore coincides with X. Thus for every $\varepsilon > 0$ there exists $n = n(\varepsilon) \geq 1$ such that $\mathrm{d}(T^n z, z) < \varepsilon$. This proves the proposition for $p = 1$.

We suppose now that $p > 1$ and the proposition holds with p replaced by $p - 1$.

(iii) We show next that, for any $\varepsilon > 0$, there exists a finite set K of positive integers such that, for all $x, x' \in X$,

$$\mathrm{d}(T^k x', x) < \varepsilon/2 \quad \text{for some } k \in K.$$

If B is a nonempty open subset of X, then for every $z \in X$ there exists some $n \geq 1$ such that $T^n z \in B$. Hence $X = \bigcup_{n\geq 1} T^{-n} B$. Since X is compact and the sets $T^{-n}B$ are open, there is a finite set $K(B)$ of positive integers such that

$$X = \bigcup_{k\in K(B)} T^{-k} B.$$

Since X is compact again, there exist finitely many open balls $B_1, \ldots, B_r$ with radius $\varepsilon/4$ such that $X = B_1 \cup \cdots \cup B_r$. If $x, x' \in X$, then $x \in B_i$ for some $i \in \{1, \ldots, r\}$ and $x' \in T^{-k} B_i$ for some $k \in K(B_i)$. Thus we can take $K = K(B_1) \cup \cdots \cup K(B_r)$.

(iv) We now show that, for any $\varepsilon > 0$ and any $x \in X$, there exists $y \in X$ and $n \geq 1$ such that

$$\mathrm{d}(T^n y, x) < \varepsilon, \quad \mathrm{d}(T^{2n} y, x) < \varepsilon, \ldots, \quad \mathrm{d}(T^{pn} y, x) < \varepsilon.$$

In fact, since each $T^k (k \in K)$ is uniformly continuous on X, we can choose $\rho > 0$ so that $\mathrm{d}(x_1, x_2) < \rho$ implies $\mathrm{d}(T^k x_1, T^k x_2) < \varepsilon/2$ for all $x_1, x_2 \in X$ and all $k \in K$. By the induction hypothesis, there exist $x' \in X$ and $n \geq 1$ such that

$$\mathrm{d}(T^n x', x') < \rho, \ldots, \mathrm{d}(T^{(p-1)n} x', x') < \rho.$$

But the invariant set TX is closed, since X is compact, and so $TX = X$. Hence $T^n X = X$ and we can choose $y' \in X$ so that $T^n y' = x'$. Thus

$$\mathrm{d}(T^n y', x') = 0, \quad \mathrm{d}(T^{2n} y', x') < \rho, \ldots, \quad \mathrm{d}(T^{pn} y', x') < \rho.$$

It follows that, for all $k \in K$,

$$\mathrm{d}(T^{n+k} y', T^k x') < \varepsilon/2, \ldots, \mathrm{d}(T^{pn+k} y', T^k x') < \varepsilon/2.$$

For each $x \in X$ there is a $k \in K$ such that $\mathrm{d}(T^k x', x) < \varepsilon/2$. Thus if $y = T^k y'$, then

$$\mathrm{d}(T^n y, x) < \varepsilon, \ldots, \mathrm{d}(T^{pn} y, x) < \varepsilon.$$

(v) Let $\varepsilon_0 > 0$ and $x_0 \in X$ be given. By (iv) there exist $x_1 \in X$ and $n_1 \geq 1$ such that

$$\mathrm{d}(T^{n_1} x_1, x_0) < \varepsilon_0, \ldots, \mathrm{d}(T^{pn_1} x_1, x_0) < \varepsilon_0.$$

We can now choose $\varepsilon_1 \in (0, \varepsilon_0)$ so that $\mathrm{d}(x, x_1) < \varepsilon_1$ implies

$$\mathrm{d}(T^{n_1} x, x_0) < \varepsilon_0, \ldots, \mathrm{d}(T^{pn_1} x, x_0) < \varepsilon_0.$$

Suppose we have defined points $x_1, \dots, x_k$, positive integers $n_1, \dots, n_k$, and $\varepsilon_1, \dots, \varepsilon_k \in (0, \varepsilon_0)$ such that, for $i = 1, \dots, k$,

$$\mathrm{d}(T^{n_i} x_i, x_{i-1}) < \varepsilon_{i-1}, \dots, \mathrm{d}(T^{pn_i} x_i, x_{i-1}) < \varepsilon_{i-1},$$

and $\mathrm{d}(x, x_i) < \varepsilon_i$ implies

$$\mathrm{d}(T^{n_i} x, x_{i-1}) < \varepsilon_{i-1}, \dots, \mathrm{d}(T^{pn_i} x, x_{i-1}) < \varepsilon_{i-1}.$$

By (iv) there exist $x_{k+1} \in X$ and $n_{k+1} \geq 1$ such that

$$\mathrm{d}(T^{n_{k+1}} x_{k+1}, x_k) < \varepsilon_k, \dots, \mathrm{d}(T^{pn_{k+1}} x_{k+1}, x_k) < \varepsilon_k,$$

and we can then choose $\varepsilon_{k+1} \in (0, \varepsilon_0)$ so that $\mathrm{d}(x, x_{k+1}) < \varepsilon_{k+1}$ implies

$$\mathrm{d}(T^{n_{k+1}} x, x_k) < \varepsilon_k, \dots, \mathrm{d}(T^{pn_{k+1}} x, x_k) < \varepsilon_k.$$

Thus the process can be continued indefinitely.

By taking successively $i = j-1, j-2, \dots$ we see that, if $i < j$, then

$$\mathrm{d}(T^{n_{i+1}+\cdots+n_{j-1}+n_j} x_j, x_i) < \varepsilon_i, \dots, \mathrm{d}(T^{p(n_{i+1}+\cdots+n_{j-1}+n_j)} x_j, x_i) < \varepsilon_i.$$

Since X is compact, it is covered by a finite number r of open balls with radius $\varepsilon_0/2$. Hence there exist i, j with $0 \leq i < j \leq r$ such that $\mathrm{d}(x_i, x_j) < \varepsilon_0$. If we put $n = n_{i+1} + \cdots + n_{j-1} + n_j$ then, since $\varepsilon_i < \varepsilon_0$, we obtain from the triangle inequality

$$\mathrm{d}(T^n x_j, x_j) < 2\varepsilon_0, \dots, \mathrm{d}(T^{pn} x_j, x_j) < 2\varepsilon_0.$$

But $\varepsilon_0 > 0$ was arbitrary. □

It may be deduced from Proposition 26, by means of *Baire's category theorem*, that under the same hypotheses there exists a point $z \in X$ and an increasing sequence (n_k) of positive integers such that $T^{in_k} z \to z$ as $k \to \infty$ $(i = 1, \dots, p)$. However, as we now show, Proposition 26 already suffices to proves van der Waerden's theorem.

The set X^* of all infinite sequences $x = (x_1, x_2, \dots)$, where $x_i \in \{1, 2, \dots, r\}$ for every $i \geq 1$, can be given the structure of a compact metric space by defining $\mathrm{d}(x, x) = 0$ and $\mathrm{d}(x, y) = 2^{-k}$ if $x \neq y$ and k is the least positive integer such that $x_k \neq y_k$. The shift map $\tau : X^* \to X^*$, defined by $\tau((x_1, x_2, \dots)) = (x_2, x_3, \dots)$, is continuous, since

$$\mathrm{d}(\tau(x), \tau(y)) \leq 2\,\mathrm{d}(x, y).$$

With the partition $\mathbb{N} = S_1 \cup \cdots \cup S_r$ in the statement of van der Waerden's theorem we associate the infinite sequence $x \in X^*$ defined by $x_i = j$ if $i \in S_j$.

Let X denote the closure of the set $(\tau^n x)_{n \geq 1}$. Then X is a closed subset of X^* which is invariant under τ. By Proposition 26, there exists a point $z \in X$ and a positive integer n such that

$$\mathrm{d}(\tau^n z, z) < 1/2, \quad \mathrm{d}(\tau^{2n} z, z) < 1/2, \dots, \quad \mathrm{d}(\tau^{pn} z, z) < 1/2;$$

i.e. $z_1 = z_{n+1} = z_{2n+1} = \cdots = z_{pn+1}$. Since $z \in X$, there is a positive integer m such that $\mathrm{d}(\tau^m x, z) < 2^{-pn-1}$, i.e. $x_{m+i} = z_i$ for $1 \le i \le pn+1$. It follows that

$$x_{m+1} = x_{m+n+1} = \cdots = x_{m+pn+1}.$$

Thus for every positive integer p there is a set $S_{j(p)}$ which contains an arithmetic progression of length p. Since there are only r possible values for $j(p)$, one of the sets S_j must contain arithmetic progressions of arbitrary finite length.

A far-reaching generalization of van der Waerden's theorem has been given by Hales and Jewett (1963). Let $A = \{a_1, \ldots, a_q\}$ be a finite set and let A^n be the set of all n-tuples with elements from A. A set $W = \{w^1, \ldots, w^q\} \subseteq A^n$ of q n-tuples $w^k = (w_1^k, \ldots, w_n^k)$ is said to be a *combinatorial line* if there exists a partition

$$\{1, \ldots, n\} = I \cup J, \ I \cap J = \emptyset,$$

such that

$$w_i^k = a_k \ (k = 1, \ldots, q) \quad \text{for } i \in I; \quad w_j^1 = \cdots = w_j^q \quad \text{for } j \in J.$$

The Hales–Jewett theorem says that, for any positive integer r, there exists a positive integer $N = N(q, r)$ such that, if A^N is partitioned into r classes, then at least one of these classes contains a combinatorial line.

If one takes $A = \{0, 1, \ldots, q-1\}$ and interprets A^n as the set of expansions to base q of all non-negative integers less than q^n, then a combinatorial line is an arithmetic progression. On the other hand, if one takes $A = \mathbb{F}_q$ to be a finite field with q elements and interprets A^n as the n-dimensional vector space $\mathbb{F}_q^n$, then a combinatorial line is an affine line. The interesting feature of the Hales–Jewett theorem is that it is purely combinatorial and does not involve any notion of addition.

6 Further Remarks

Uniform distribution and discrepancy are thoroughly discussed in Kuipers and Niederreiter [30]. For later results, see Drmota and Tichy [13]. Since these two books have extensive bibliographies, we will be sparing with references. However, it would be remiss not to recommend the great paper of Weyl [52], which remains as fresh as when it was written.

Lemma 0 is often attributed to Polya (1920), but it was already proved by Buchanan and Hildebrandt [9].

Fejér's proof that continuous periodic functions can be uniformly approximated by trigonometric polynomials is given in Dym and McKean [15]. The theorem also follows directly from the the theorem of Weierstrass (1885) on the uniform approximation of continuous functions by ordinary polynomials. A remarkable generalization of both results was given by Stone (1937); see Stone [49]. The 'Stone–Weierstrass theorem' is also proved in Rudin [44], for example.

Chen [11] gives a quantitative version of Kronecker's theorem of a different type from Proposition 3′.

The converse of Proposition 10 is proved by Kemperman [27]. For the history of the problem of mean motion, and generalizations to almost periodic functions, see Jessen and Tornehave [24]. Methods for estimating exponential sums were developed in connection with the theory of uniform distribution, but then found other applications. See Chandrasekharan [10] and Graham and Kolesnik [21].

For applications of discrepancy to numerical integration, see Niederreiter [36, 37]. For the basic properties of functions of bounded variation and the definition of total variation see, for example, Riesz and Sz.-Nagy [42].

Sharper versions of the original Erdős–Turan inequality are proved by Niederreiter and Philipp [38] and in Montgomery [35]. The discrepancy of the sequence $(\{n\alpha\})$, where α is an irrational number whose continued fraction expansion has bounded partial quotients (i.e., is *badly approximable*), is discussed by Dupain and Sós [14]. The discrepancy of the sequence $(\{n\alpha\})$, where $\alpha \in \mathbb{R}^d$, has been deeply studied by Beck [3]. The work of Roth, Schmidt and others is treated in Beck and Chen [4].

For accounts of measure theory, see Billingsley [6], Halmos [22], Loève [32] and Saks [46]. More detailed treatments of ergodic theory are given in the books of Petersen [39], Walters [51] and Cornfeld *et al.* [12]. The prehistory of ergodic theory is described by the Ehrenfests [16]. However, they do not refer to the paper of Poincaré (1894), which is reproduced in [41].

The proof of Birkhoff's ergodic theorem given here follows Katznelson and Weiss [26]. A different proof is given in the book of Walters.

Many other ergodic theorems besides Birkhoff's are discussed in Krengel [29]. We mention only the *subadditive ergodic theorem* of Kingman (1968): if T is a measure-preserving transformation of the probability space $(X, \mathscr{B}, \mu)$ and if (g_n) is a sequence of functions in $L(X, \mathscr{B}, \mu)$ such that $\inf_n n^{-1} \int_X g_n \, \mathrm{d}\mu > -\infty$ and, for all $m, n \geq 1$,

$$g_{n+m}(x) \leq g_n(x) + g_m(T^n x) \text{ a.e.},$$

then $n^{-1} g_n(x) \to g^*(x)$ a.e., where $g^*(Tx) = g^*(x)$ a.e., $g^* \in L(X, \mathscr{B}, \mu)$ and

$$\int_X g^* \mathrm{d}\mu = \lim_{n\to\infty} n^{-1} \int_X g_n \, \mathrm{d}\mu = \inf_n n^{-1} \int_X g_n \, \mathrm{d}\mu.$$

Birkhoff's ergodic theorem may be regarded as a special case by taking $g_n(x) = \sum_{k=0}^{n-1} f(T^k x)$. A simple proof of Kingman's theorem is given by Steele [48]. For applications of Kingman's theorem to percolation processes and products of random matrices, see Kingman [28]. The multiplicative ergodic theorem of Oseledets is derived from Kingman's theorem by Ruelle [45].

The book of Kuipers and Niederreiter cited above has an extensive discussion of normal numbers. For normality with respect to a matrix, see also Brown and Moran [8].

Proofs of Gauss's statement on the continued fraction map are contained in the books by Billingsley [7] and Rockett and Szusz [43]. For more recent work, see Wirsing [53], Babenko [2] and Mayer [33]. For the deviation of $(1/n) \log q_n(\xi)$ from its (a.e.) limiting value $\pi^2/(12 \log 2)$ there are analogues of the central limit theorem and the law of the iterated logarithm; see Philipp and Stackelberg [40]. For higher-dimensional generalizations of Gauss's invariant measure, see Hardcastle and Khanin [23].

Applications of ergodic theory to classical mechanics are discussed in the books of Arnold and Avez [1] and Katok and Hasselblatt [25]. For connections between ergodic theory and the '$3x + 1$ problem', see Lagarias [31].

Ergodic theory has been used to generalize considerably some of the results on lattices in Chapter VIII. A *lattice* in a locally compact group G is a discrete subgroup Γ such that the G-invariant measure of the quotient space G/Γ is finite. (In Chapter VIII, $G = \mathbb{R}^n$ and $\Gamma = \mathbb{Z}^n$.) Zimmer [54] gives a good introduction to the results which have been obtained in this area.

An attractive account of the work of Furstenberg and his collaborators is given in Furstenberg [17]. See also Graham *et al.* [20] and the book of Petersen cited above. The discovery of van der Waerden's theorem is described in van der Waerden [50]. For a recent direct proof, see Mills [34].

The direct proofs reduce the theorem to an equivalent finite form: *for any positive integer p, there exists a positive integer N such that, whenever the set $\{1, 2, \ldots, N\}$ is partitioned into two subsets, at least one subset contains an arithmetic progression of length p*. The original proofs provided an upper bound for the least possible value $N(p)$ of N, but it was unreasonably large. Some progress towards obtaining reasonable upper bounds has recently been made by Shelah [47] and Gowers [19].

The Hales–Jewett theorem is proved, and then extensively generalized, in Bergelson and Leibman [5]. Furstenberg and Katznelson [18] prove a density version of the Hales–Jewett theorem, analogous to Szemeredi's density version of van der Waerden's theorem.

7 Selected References

[1] V.I. Arnold and A. Avez, *Ergodic problems of classical mechanics*, Benjamin, New York, 1968.

[2] K.I. Babenko, On a problem of Gauss, *Soviet Math. Dokl.* **19** (1978), 136–140.

[3] J. Beck, Probabilistic diophantine approximation, I. Kronecker sequences, *Ann. of Math.* **140** (1994), 451–502.

[4] J. Beck and W.W.L. Chen, *Irregularities of distribution*, Cambridge University Press, 1987.

[5] V. Bergelson and A. Leibman, Set polynomials and polynomial extension of the Hales–Jewett theorem, *Ann. of Math.* **150** (1999), 33–75.

[6] P. Billingsley, *Probability and measure*, 3rd ed., Wiley, New York, 1995.

[7] P. Billingsley, *Ergodic theory and information*, reprinted, Krieger, Huntington, N.Y., 1978.

[8] G. Brown and W. Moran, Schmidt's conjecture on normality for commuting matrices, *Invent. Math.* **111** (1993), 449–463.

[9] H.E. Buchanan and H.T. Hildebrandt, Note on the convergence of a sequence of functions of a certain type, *Ann. of Math.* **9** (1908), 123–126.

[10] K. Chandrasekharan, Exponential sums in the development of number theory, *Proc. Steklov Inst. Math.* **132** (1973), 3–24.

[11] Y.-G. Chen, The best quantitative Kronecker's theorem, *J. London Math. Soc.* (2) **61** (2000), 691–705.

[12] I.P. Cornfeld, S.V. Fomin and Ya. G. Sinai, *Ergodic theory*, Springer-Verlag, New York, 1982.

[13] M. Drmota and R.F. Tichy, *Sequences, discrepancies and applications*, Lecture Notes in Mathematics **1651**, Springer, Berlin, 1997.

[14] I. Dupain and V.T. Sós, On the discrepancy of $(n\alpha)$ sequences, *Topics in classical number theory* (ed. G. Halász), Vol. I, pp. 355–387, North-Holland, Amsterdam, 1984.
[15] H. Dym and H.P. McKean, *Fourier series and integrals*, Academic Press, Orlando, FL, 1972.
[16] P. and T. Ehrenfest, *The conceptual foundations of the statistical approach in mechanics*, English translation by M.J. Moravcsik, Cornell University Press, Ithaca, 1959. [German original, 1912]
[17] H. Furstenberg, *Recurrence in ergodic theory and combinatorial number theory*, Princeton University Press, 1981.
[18] H. Furstenberg and Y. Katznelson, A density version of the Hales–Jewett theorem, *J. Analyse Math.* **57** (1991), 64–119.
[19] W.T. Gowers, A new proof of Szemeredi's theorem, *Geom. Funct. Anal.* **11** (2001), 465–588.
[20] R.L. Graham, B.L. Rothschild and J.H. Spencer, *Ramsey theory*, 2nd ed., Wiley, New York, 1990.
[21] S.W. Graham and G. Kolesnik, *Van der Corput's method of exponential sums*, London Math. Soc. Lecture Notes **126**, Cambridge University Press, 1991.
[22] P.R. Halmos, *Measure theory*, 2nd printing, Springer-Verlag, New York, 1974.
[23] D.M. Hardcastle and K. Khanin, Continued fractions and the d-dimensional Gauss transformation, *Comm. Math. Phys.* **215** (2001), 487–515.
[24] B. Jessen and H. Tornehave, Mean motion and zeros of almost periodic functions, *Acta Math.* **77** (1945), 137–279.
[25] A. Katok and B. Hasselblatt, *Introduction to the modern theory of dynamical systems*, Cambridge University Press, 1995.
[26] Y. Katznelson and B. Weiss, A simple proof of some ergodic theorems, *Israel J. Math.* **42** (1982), 291–296.
[27] J.H.B. Kemperman, Distributions modulo 1 of slowly changing sequences, *Nieuw Arch. Wisk.* (3) **21** (1973), 138–163.
[28] J.F.C. Kingman, Subadditive processes, *Ecole d'Eté de Probabilités de Saint-Flour* V-1975 (ed. A. Badrikian), pp. 167–223, Lecture Notes in Mathematics **539**, Springer-Verlag, 1976.
[29] U. Krengel, *Ergodic theorems*, de Gruyter, Berlin, 1985.
[30] L. Kuipers and H. Niederreiter, *Uniform distribution of sequences*, Wiley, New York, 1974.
[31] J.C. Lagarias, The $3x + 1$ problem and its generalizations, *Amer. Math. Monthly* **92** (1985), 3–23.
[32] M. Loève, *Probability theory*, 4th ed. in 2 vols., Springer-Verlag, New York, 1978.
[33] D.H. Mayer, On the thermodynamic formalism for the Gauss map, *Comm. Math. Phys.* **130** (1990), 311–333.
[34] G. Mills, A quintessential proof of van der Waerden's theorem on arithmetic progressions, *Discrete Math.* **47** (1983), 117–120.
[35] H.L. Montgomery, *Ten lectures on the interface between analytic number theory and harmonic analysis*, CBMS Regional Conference Series in Mathematics **84**, American Mathematical Society, Providence, R.I., 1994.
[36] H. Niederreiter, Quasi-Monte Carlo methods and pseudo-random numbers, *Bull. Amer. Math. Soc.* **84** (1978), 957–1041.
[37] H. Niederreiter, *Random number generation and quasi-Monte Carlo methods*, CBMS–NSF Regional Conference Series in Applied Mathematics **63**, SIAM, Philadelphia, 1992.
[38] H. Niederreiter and W. Philipp, Berry–Esseen bounds and a theorem of Erdös and Turán on uniform distribution mod 1, *Duke Math. J.* **40** (1973), 633–649.
[39] K. Petersen, *Ergodic theory*, Cambridge University Press, 1983.
[40] W. Philipp and O.P. Stackelberg, Zwei Grenzwertsätze für Kettenbrüche, *Math. Ann.* **181** (1969), 152–156.

[41] H. Poincaré, Sur la théorie cinétique des gaz, *Oeuvres*, t. X, pp. 246–263, Gauthier-Villars, Paris, 1954.
[42] F. Riesz and B. Sz.-Nagy, *Functional analysis*, English transl. by L.F. Boron, Ungar, New York, 1955.
[43] A. Rockett and P. Szusz, *Continued fractions*, World Scientific, Singapore, 1992.
[44] W. Rudin, *Principles of mathematical analysis*, 3rd ed., McGraw-Hill, New York, 1976.
[45] D. Ruelle, Ergodic theory of differentiable dynamical systems, *Inst. Hautes Études Sci. Publ. Math.* **50** (1979), 27–58.
[46] S. Saks, *Theory of the integral*, 2nd revised ed., English transl. by L.C. Young, reprinted, Dover, New York, 1964.
[47] S. Shelah, Primitive recursion bounds for van der Waerden numbers, *J. Amer. Math. Soc.* **1** (1988), 683–697.
[48] J.M. Steele, Kingman's subadditive ergodic theorem, *Ann. Inst. H. Poincaré Sect. B* **25** (1989), 93–98.
[49] M.H. Stone, A generalized Weierstrass approximation theorem, *Studies in modern analysis* (ed. R.C. Buck), pp. 30–87, Mathematical Association of America, 1962.
[50] B.L. van der Waerden, How the proof of Baudet's conjecture was found, *Studies in Pure Mathematics* (ed. L. Mirsky), pp. 251–260, Academic Press, London, 1971.
[51] P. Walters, *An introduction to ergodic theory*, Springer-Verlag, New York, 1982.
[52] H. Weyl, Über die Gleichverteilung von Zahlen mod Eins, *Math. Ann.* **77** (1916), 313–352. [Reprinted in *Selecta Hermann Weyl*, pp. 111–147, Birkhäuser, Basel, 1956 and in *Hermann Weyl, Gesammelte Abhandlungen* (ed. K. Chandrasekharan), *Band I*, pp. 563–599, Springer-Verlag, Berlin, 1968]
[53] E. Wirsing, On the theorem of Gauss–Kusmin–Lévy and a Frobenius type theorem for function spaces, *Acta Arith.* **24** (1974), 507–528.
[54] R.J. Zimmer, *Ergodic theory and semi-simple groups*, Birkhäuser, Boston, 1984.

Additional Reference

B. Kra, The Green-Tao theorem on arithmetic progressions in the primes: an ergodic point of view, *Bull. Amer. Math. Soc.* (*N.S.*) **43** (2006), 3–23.

XII

Elliptic Functions

Our discussion of elliptic functions may be regarded as an essay in revisionism, since we do not use Liouville's theorem, Riemann surfaces or the Weierstrassian functions. We wish to show that the methods used by the founding fathers of the subject provide a natural and rigorous approach, which is very well suited for applications.

The work is arranged so that the initial sections are mutually independent, although motivation for each section is provided by those which precede it. To some extent we have also separated the discussion for real and for complex parameters, so that those interested only in the real case may skip the complex one.

1 Elliptic Integrals

After the development of the integral calculus in the second half of the 17th century, it was natural to apply it to the determination of the arc length of an ellipse since, by Kepler's first law, the planets move in elliptical orbits with the sun at one focus.

An ellipse is described in rectangular coordinates by an equation

$$x^2/a^2 + y^2/b^2 = 1,$$

where a and b are the *semi-axes* of the ellipse ($a > b > 0$). It is also given parametrically by

$$x = a\sin\theta,\ \ y = b\cos\theta \quad (0 \le \theta \le 2\pi).$$

The arc length $s(\Theta)$ from $\theta = 0$ to $\theta = \Theta$ is given by

$$\begin{aligned} s(\Theta) &= \int_0^{\Theta} [(dx/d\theta)^2 + (dy/d\theta)^2]^{1/2} d\theta \\ &= \int_0^{\Theta} (a^2\cos^2\theta + b^2\sin^2\theta)^{1/2} d\theta \\ &= \int_0^{\Theta} [a^2 - (a^2 - b^2)\sin^2\theta]^{1/2} d\theta. \end{aligned}$$

W.A. Coppel, *Number Theory: An Introduction to Mathematics*, Universitext,
DOI: 10.1007/978-0-387-89486-7_12, © Springer Science + Business Media, LLC 2009

If we put $b^2 = a^2(1 - k^2)$, where k $(0 < k < 1)$ is the *eccentricity* of the ellipse, this takes the form

$$s(\Theta) = a \int_0^{\Theta} (1 - k^2 \sin^2 \theta)^{1/2} d\theta.$$

If we further put $z = \sin\theta = x/a$ and restrict attention to the first quadrant, this assumes the algebraic form

$$a \int_0^{Z} [(1 - k^2 z^2)/(1 - z^2)]^{1/2} dz.$$

Since the arc length of the whole quadrant is obtained by taking $Z = 1$, the arc length of the whole ellipse is

$$L = 4a \int_0^{1} [(1 - k^2 z^2)/(1 - z^2)]^{1/2} dz.$$

Consider next Galileo's problem of the simple pendulum. If θ is the angle of deflection from the downward vertical, the equation of motion of the pendulum is

$$d^2\theta/dt^2 + (g/l) \sin\theta = 0,$$

where l is the length of the pendulum and g is the gravitational constant. This differential equation has the first integral

$$(d\theta/dt)^2 = (2g/l)(\cos\theta - a),$$

where a is a constant. In fact $a < 1$ for a real motion, and for oscillatory motion we must also have $a > -1$. We can then put $a = \cos\alpha$ $(0 < \alpha < \pi)$, where α is the maximum value of θ, and integrate again to obtain

$$\begin{aligned} t &= (l/2g)^{1/2} \int_0^{\Theta} (\cos\theta - \cos\alpha)^{-1/2} d\theta \\ &= (l/4g)^{1/2} \int_0^{\Theta} (\sin^2 \alpha/2 - \sin^2 \theta/2)^{-1/2} d\theta. \end{aligned}$$

Putting $k = \sin\alpha/2$ and $kx = \sin\theta/2$, we can rewrite this in the form

$$t = (l/g)^{1/2} \int_0^{X} [(1 - k^2 x^2)(1 - x^2)]^{-1/2} dx.$$

The angle of deflection θ attains its maximum value α when $X = 1$, and the motion is periodic with period

$$T = 4(l/g)^{1/2} \int_0^{1} [(1 - k^2 x^2)(1 - x^2)]^{-1/2} dx.$$

Attempts to evaluate the integrals in both these problems in terms of algebraic and elementary transcendental functions proved fruitless. Thus the idea arose of treating them as fundamental entities in terms of which other integrals could be expressed.

An example is the determination of the arc length of a *lemniscate*. This curve, which was studied by Jacob Bernoulli (1694), has the form of a figure of eight and is the locus of all points $z \in \mathbb{C}$ such that $|2z^2 - 1| = 1$ or, in polar coordinates,

$$r^2 = \cos 2\theta \quad (-\pi/4 \le \theta \le \pi/4 \cup 3\pi/4 \le \theta \le 5\pi/4).$$

If $-\pi/4 \le \Theta \le 0$, the arc length $s(\Theta)$ from $\theta = -\pi/4$ to $\theta = \Theta$ is given by

$$\begin{aligned} s(\Theta) &= \int_{-\pi/4}^{\Theta} [r^2 + (dr/d\theta)^2]^{1/2} d\theta \\ &= \int_{-\pi/4}^{\Theta} [r^2 + (1 - r^4)/r^2]^{1/2} d\theta \\ &= \int_0^R (1 - r^4)^{-1/2} dr. \end{aligned}$$

If we make the change of variables $x = \sqrt{2}r/(1 + r^2)^{1/2}$, then on account of $dx/dr = \sqrt{2}/(1 + r^2)^{3/2}$ we obtain

$$s(\Theta) = 2^{-1/2} \int_0^X [(1 - x^2/2)(1 - x^2)]^{-1/2} dx.$$

Another example is the determination of the surface area of an ellipsoid. Suppose the ellipsoid is described in rectangular coordinates by the equation

$$x^2/a^2 + y^2/b^2 + z^2/c^2 = 1,$$

where $a > b > c > 0$. The total surface area is $8S$, where S is the surface area of the part contained in the positive octant. In this octant we have

$$z = c[1 - (x/a)^2 - (y/b)^2]^{1/2}$$

and hence

$$1 + (\partial z/\partial x)^2 + (\partial z/\partial y)^2 = [1 - (\alpha x/a)^2 - (\beta y/b)^2]/[1 - (x/a)^2 - (y/b)^2],$$

where

$$\alpha = (a^2 - c^2)^{1/2}/a, \quad \beta = (b^2 - c^2)^{1/2}/b.$$

Consequently

$$S = \int_0^a \int_0^{b(1-(x/a)^2)^{1/2}} [1 - (\alpha x/a)^2 - (\beta y/b)^2]^{1/2} [1 - (x/a)^2 - (y/b)^2]^{-1/2} dy dx.$$

If we make the change of variables

$$x = ar\cos\theta, \quad y = br\sin\theta,$$

with Jacobian $J = abr$, we obtain

$$S = ab \int_0^{\pi/2} d\theta \int_0^1 (1 - \sigma r^2)^{1/2} (1 - r^2)^{-1/2} r dr,$$

where

$$\sigma = \alpha^2 \cos^2\theta + \beta^2 \sin^2\theta.$$

If we now put

$$u^2 = (1 - r^2)/(1 - \sigma r^2),$$

then $r^2 = (1 - u^2)/(1 - \sigma u^2)$ and

$$r\,dr/du = -(1 - \sigma)u/(1 - \sigma u^2)^2.$$

Hence

$$S = ab \int_0^{\pi/2} d\theta \int_0^1 (1 - \sigma)(1 - \sigma u^2)^{-2} du.$$

Inverting the order of integration and giving σ its value, we obtain

$$S = ab \int_0^1 du \int_0^{\pi/2} [(1 - \alpha^2)\cos^2\theta + (1 - \beta^2)\sin^2\theta] \times [(1 - \alpha^2u^2)\cos^2\theta + (1 - \beta^2u^2)\sin^2\theta]^{-2} d\theta.$$

It is readily verified that

$$\int_0^{\pi/2} \cos^2\theta (m\cos^2\theta + n\sin^2\theta)^{-2} d\theta = \pi/4m(mn)^{1/2},$$
$$\int_0^{\pi/2} \sin^2\theta (m\cos^2\theta + n\sin^2\theta)^{-2} d\theta = \pi/4n(mn)^{1/2}.$$

Thus we obtain finally

$$S = (\pi ab/4) \int_0^1 [(1 - \alpha^2)/(1 - \alpha^2u^2) + (1 - \beta^2)/(1 - \beta^2u^2)] \times [(1 - \alpha^2u^2)(1 - \beta^2u^2)]^{-1/2} du.$$

By an *elliptic integral* one understands today any integral of the form

$$\int R(x, w)\, dx,$$

where $R(x, w)$ is a rational function of x and w, and where $w^2 = g(x)$ is a polynomial in x of degree 3 or 4 without repeated roots. The elliptic integral is said to be *complete* if it is a definite integral in which the limits of integration are distinct roots of $g(x)$.

The case of a quartic is easily reduced to that of a cubic. In the preceding examples we can simply put $y = x^2$. Thus, for the lemniscate,

$$s(\Theta) = 2^{-1/2} \int_0^Y [4y(1 - y)(1 - y/2)]^{-1/2} dy.$$

In general, suppose $g(x) = (x - \alpha)h(x)$, where h is a cubic. If

$$h(x) = h_0(x-\alpha)^3 + h_1(x-\alpha)^2 + h_2(x-\alpha) + h_3$$

and we make the change of variables $x = \alpha + 1/y$, then $g(x) = g^*(y)/y^4$, where

$$g^*(y) = h_0 + h_1 y + h_2 y^2 + h_3 y^3,$$

and

$$\int R(x, w)\, dx = \int R^*(y, v)\, dy,$$

where $R^*(y, v)$ is a rational function of y and v, and $v^2 = g^*(y)$.

Since any even power of w is a polynomial in x, the integrand can be written in the form $R(x, w) = (A + Bw)/(C + Dw)$, where A, B, C, D are polynomials in x. Multiplying numerator and denominator by $(C - Dw)w$, we obtain

$$R(x, w) = N/L + M/Lw,$$

where L, M, N are polynomials in x. By decomposing the rational function N/L into partial fractions its integral can be evaluated in terms of rational functions and (real or complex) logarithms. By similarly decomposing the rational function M/L into partial fractions, we are reduced to evaluating the integrals

$$I_0 = \int dx/w, \quad I_n = \int x^n dx/w, \quad J_n(\gamma) = \int (x-\gamma)^{-n} dx/w,$$

where $n \in \mathbb{N}$ and $\gamma \in \mathbb{C}$.

The argument of the preceding paragraph is actually valid if $w^2 = g$ is any polynomial. Suppose now that g is a cubic without repeated roots, say

$$g(x) = a_0x^3 + a_1x^2 + a_2x + a_3.$$

By differentiation we obtain, for any integer $m \geq 0$,

$$(x^m w)' = mx^{m-1}w + x^m g'/2w = (2mx^{m-1}g + x^m g')/2w.$$

Since the numerator on the right is the polynomial

$$(2m+3)a_0x^{m+2} + (2m+2)a_1x^{m+1} + (2m+1)a_2x^m + 2ma_3x^{m-1},$$

it follows on integration that

$$2x^m w = (2m+3)a_0I_{m+2} + (2m+2)a_1I_{m+1} + (2m+1)a_2I_m + 2ma_3I_{m-1}.$$

It follows by induction that, for each integer $n > 1$,

$$I_n = p_n(x)w + c_nI_0 + c_n'I_1,$$

where $p_n(x)$ is a polynomial of degree $n-2$ and c_n, c_n' are constants. Thus the evaluation of I_n for $n > 1$ reduces to the evaluation of I_0 and I_1.

Consider now the integral $J_n(\gamma)$. In the same way as before, for any integer $m \geq 1$,

$$\begin{aligned} d\{(x-\gamma)^{-m}w\}/dx &= -m(x-\gamma)^{-m-1}w + (x-\gamma)^{-m}g'/2w \\ &= \{-2mg + (x-\gamma)g'\}/2w(x-\gamma)^{m+1}. \end{aligned}$$

We can write

$$g(x) = b_0 + b_1(x-\gamma) + b_2(x-\gamma)^2 + b_3(x-\gamma)^3$$

and the numerator on the right of the previous equation is then

$$-2mb_0 + (1-2m)b_1(x-\gamma) + (2-2m)b_2(x-\gamma)^2 + (3-2m)b_3(x-\gamma)^3.$$

It follows on integration that

$$\begin{aligned} 2(x-\gamma)^{-m}w = &-2mb_0J_{m+1}(\gamma) + (1-2m)b_1J_m(\gamma) \\ &+ (2-2m)b_2J_{m-1}(\gamma) + (3-2m)b_3J_{m-2}(\gamma), \end{aligned}$$

where $J_{-1}(\gamma) = \int (x-\gamma)\,dx/w$ is a constant linear combination of I_0 and I_1. Since g does not have repeated roots, $b_1 \neq 0$ if $b_0 = 0$.

It follows by induction that if $g(\gamma) = b_0 \neq 0$ then, for any $n > 1$,

$$J_n(\gamma) = q_n((x-\gamma)^{-1})w + d_nJ_1(\gamma) + d_n'I_0 + d_n''I_1,$$

where $q_n(t)$ is a polynomial of degree $n-1$ and d_n, d_n', d_n'' are constants. On the other hand, if $g(\gamma) = 0$ then $g'(\gamma) = b_1 \neq 0$ and, for any $n \geq 1$,

$$J_n(\gamma) = r_n((x-\gamma)^{-1})w + e_nI_0 + e_n'I_1,$$

where $r_n(t)$ is a polynomial of degree n and e_n, e_n' are constants.

Thus the evaluation of an arbitrary elliptic integral can be reduced to the evaluation of

$$I_0 = \int dx/w, \quad I_1 = \int x\,dx/w, \quad J_1(\gamma) = \int (x-\gamma)^{-1}dx/w,$$

where $w^2 = g$ is a cubic without repeated roots, $\gamma \in \mathbb{C}$ and $g(\gamma) \neq 0$. Following Legendre (1793), to whom this reduction is due, integrals of these types are called respectively *elliptic integrals of the first, second and third kinds*.

The cubic g can itself be simplified. If α is a root of g then, by replacing x by $x-\alpha$, we may assume that $g(0) = 0$. If β is now another root of g then, by replacing x by x/β, we may further assume that $g(1) = 0$. Thus the evaluation of an arbitrary elliptic integral may be reduced to one for which g has the form

$$g_\lambda(x) := 4x(1-x)(1-\lambda x),$$

where $\lambda \in \mathbb{C}$ and $\lambda \neq 0, 1$. This normal form, which was used by Riemann (1858) in lectures, is obtained from the normal form of Legendre by the change of variables $x = \sin^2\theta$. To draw attention to the difference, it is convenient to call it *Riemann's normal form*.

The range of λ can be further restricted by linear changes of variables. The transformation $y = (1 - \lambda x)/(1 - \lambda)$ replaces Riemann's normal form by one of the same type with λ replaced by $U\lambda = 1 - \lambda$. Similarly, the transformation $y = 1 - \lambda x$ replaces Riemann's normal form by one of the same type with λ replaced by $V\lambda = 1/(1 - \lambda)$. The transformations U and V together generate a group $\mathscr{G}$ of order 6 (isomorphic to the symmetric group $\mathscr{S}_3$ of all permutations of three letters), since

$$U^2 = V^3 = (UV)^2 = I.$$

The values of λ corresponding to the elements I, V, V^2, U, UV, UV^2 of $\mathscr{G}$ are

$$\lambda, \quad 1/(1-\lambda), \quad (\lambda-1)/\lambda, \quad 1-\lambda, \quad \lambda/(\lambda-1), \quad 1/\lambda.$$

The region $\mathscr{F}$ of the complex plane $\mathbb{C}$ defined by the inequalities

$$|\lambda - 1| < 1, \quad 0 < \mathscr{R}\lambda < 1/2,$$

is a *fundamental domain* for the group $\mathscr{G}$; i.e., no point of $\mathscr{F}$ is mapped to a different point of $\mathscr{F}$ by an element of $\mathscr{G}$ and each point of $\mathbb{C}$ is mapped to a point of $\mathscr{F}$ or its boundary $\partial\mathscr{F}$ by some element of $\mathscr{G}$. Consequently the sets $\{G(\mathscr{F}) : G \in \mathscr{G}\}$ form a *tiling* of $\mathbb{C}$; i.e.,

$$\mathbb{C} = \bigcup_{G\in\mathscr{G}} G(\mathscr{F} \cup \partial\mathscr{F}), \quad G(\mathscr{F}) \cap G'(\mathscr{F}) = \emptyset \quad \text{if } G, G' \in \mathscr{G} \text{ and } G \neq G'.$$

This is illustrated in Figure 1, where the set $G(\mathscr{F})$ is represented simply by the group element G and, in particular, $\mathscr{F}$ is represented by I. It follows that in Riemann's normal form we may suppose $\lambda \in \mathscr{F} \cup \partial\mathscr{F}$.

The changes of variable in the preceding reduction to Riemann's normal form may be complex, even though the original integrand was real. It will now be shown that any real elliptic integral can be reduced by a real change of variables to one in Riemann's normal form, where $0 < \lambda < 1$ and the independent variable is restricted to the interval $0 \le x \le 1$.

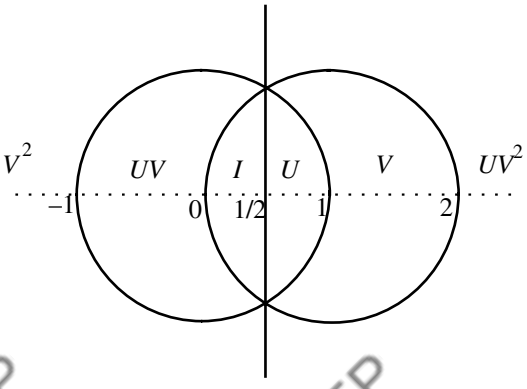

Fig. 1. Fundamental domain for λ.

If g is a cubic or quartic with only real roots, this can be achieved by a linear fractional transformation, mapping roots of g to roots of g_λ. Appropriate transformations are listed in Tables 1 and 2. It should be noted that λ is always a *cross-ratio* of the four roots of g in Table 2, and that λ is always a cross-ratio of the three roots of g and the point '∞' in Table 1.

Table 1. Reduction to Riemann's normal form, g a cubic with all roots real

$$dx/g(x)^{1/2} = dy/\mu g_\lambda(y)^{1/2}$$

$$g(x) = A(x-\alpha_1)(x-\alpha_2)(x-\alpha_3), \quad \text{where } \alpha_1 > \alpha_2 > \alpha_3; \ \ \alpha_{jk} = \alpha_j - \alpha_k$$

$$g_\lambda(y) = 4y(1-y)(1-\lambda y), \quad \text{where } 0 < \lambda < 1, \quad y \in (0,1)$$

$$\mu = (\alpha_{13})^{1/2}/2, \quad \lambda_0 = \alpha_{23}/\alpha_{13}, \quad 1-\lambda_0 = \alpha_{12}/\alpha_{13}.$$

A	λ	*Range*	*Transformation*	*Corresponding values*	
$+1$	λ_0	$x \geq \alpha_1$	$y = (x-\alpha_1)/(x-\alpha_2)$	$x = \infty$	$y = 1$
				α_1	0
-1	$1-\lambda_0$	$\alpha_2 \leq x \leq \alpha_1$	$= \alpha_{13}(x-\alpha_2)/\alpha_{12}(x-\alpha_3)$	α_1	1
				α_2	0
$+1$	λ_0	$\alpha_3 \leq x \leq \alpha_2$	$= (x-\alpha_3)/\alpha_{23}$	α_2	1
				α_3	0
-1	$1-\lambda_0$	$x \leq \alpha_3$	$= \alpha_{13}/(\alpha_1 - x)$	α_3	1
				$-\infty$	0

Table 2. Reduction to Riemann's normal form, g a quartic with all roots real

$$dx/g(x)^{1/2} = dy/\mu g_\lambda(y)^{1/2}$$

$$g(x) = A(x-\alpha_1)(x-\alpha_2)(x-\alpha_3)(x-\alpha_4), \ \ \text{where } \alpha_1 > \alpha_2 > \alpha_3 > \alpha_4; \ \alpha_{jk} = \alpha_j - \alpha_k$$

$$g_\lambda(y) = 4y(1-y)(1-\lambda y), \quad \text{where } 0 < \lambda < 1, \quad y \in (0,1)$$

$$\mu = (\alpha_{13}\alpha_{24})^{1/2}/2, \quad \lambda_0 = \alpha_{23}\alpha_{14}/\alpha_{13}\alpha_{24}, \quad 1-\lambda_0 = \alpha_{12}\alpha_{34}/\alpha_{13}\alpha_{24}.$$

A	λ	*Range*	*Transformation*	*Corresponding values*	
$+1$	λ_0	$x \geq \alpha_1$	$y = \alpha_{24}(x-\alpha_1)/\alpha_{14}(x-\alpha_2)$	$x = \infty$	$y = \alpha_{24}/\alpha_{14}$
				α_1	0
-1	$1-\lambda_0$	$\alpha_2 \leq x \leq \alpha_1$	$= \alpha_{13}(x-\alpha_2)/\alpha_{12}(x-\alpha_3)$	α_1	1
				α_2	0
$+1$	λ_0	$\alpha_3 \leq x \leq \alpha_2$	$= \alpha_{24}(x-\alpha_3)/\alpha_{23}(x-\alpha_4)$	α_2	1
				α_3	0
-1	$1-\lambda_0$	$\alpha_4 \leq x \leq \alpha_3$	$= \alpha_{13}(x-\alpha_4)/\alpha_{34}(\alpha_1 - x)$	α_3	1
				α_4	0
$+1$	λ_0	$x \leq \alpha_4$	$= \alpha_{24}(x-\alpha_1)/\alpha_{14}(x-\alpha_2)$	α_4	1
				$-\infty$	α_{24}/α_{14}

Suppose now that g is a real cubic or quartic with a pair of conjugate complex roots. Then we can write

$$g(x) = Q_1 Q_2 = (a_1 x^2 + 2b_1 x + c_1)(a_2 x^2 + 2b_2 x + c_2),$$

where the coefficients are real, $a_1 c_1 - b_1^2 > 0$ and $a_2 c_2 - b_2^2 \neq 0$, but a_2 may be zero.

Consider first the case where $a_2 \neq 0$ and $b_1 = b_2 a_1/a_2$. Then

$$Q_1 = a_1(x + b_1/a_1)^2 + b_1', \quad Q_2 = a_2(x + b_1/a_1)^2 + b_2',$$

where

$$b_1' = (a_1 c_1 - b_1^2)/a_1, \quad b_2' = (a_2 c_2 - b_2^2)/a_2.$$

If we put $y = (x + b_1/a_1)^2$, then

$$R(x) = R_1(y) + R_2(y) y^{1/2},$$

where the rational functions R_1, R_2 are determined by the rational function R, and

$$dx/g(x)^{1/2} = \pm dy/2[y(a_1 y + b_1')(a_2 y + b_2')]^{1/2}.$$

Thus we are reduced to the case of a cubic with 3 distinct real roots.

In the remaining cases there exist distinct real values s_1, s_2 of s such that the polynomial $Q_1 + sQ_2$ is proportional to a perfect square. For $Q_1 + sQ_2$ is proportional to a perfect square if

$$D(s) := (a_1 + sa_2)(c_1 + sc_2) - (b_1 + sb_2)^2 = 0.$$

We have $D(0) = a_1 c_1 - b_1^2 > 0$. If $a_2 = 0$, then $b_2 \neq 0$ and $D(\pm\infty) = -\infty$. On the other hand, if $a_2 \neq 0$, then $D(-a_1/a_2) < 0$, since $b_1 \neq b_2 a_1/a_2$, and $D(s)$ has the sign of $a_2 c_2 - b_2^2$ for both large positive and large negative s. Thus the quadratic $D(s)$ has distinct real roots s_1, s_2. Hence

$$Q_1 + s_1 Q_2 = (a_1 + s_1 a_2)(x + d_1)^2, \quad Q_1 + s_2 Q_2 = (a_1 + s_2 a_2)(x + d_2)^2,$$

where $a_1 + s_j a_2 \neq 0$ $(j = 1, 2)$ and

$$d_1 = (b_1 + s_1 b_2)/(a_1 + s_1 a_2), \quad d_2 = (b_1 + s_2 b_2)/(a_1 + s_2 a_2).$$

Consequently

$$Q_1 = A_1(x + d_1)^2 + B_1(x + d_2)^2, \quad Q_2 = A_2(x + d_1)^2 + B_2(x + d_2)^2,$$

where

$$A_1 = -s_2(a_1 + s_1 a_2)/(s_1 - s_2), \quad B_1 = s_1(a_1 + s_2 a_2)/(s_1 - s_2),$$
$$A_2 = (a_1 + s_1 a_2)/(s_1 - s_2), \quad B_2 = -(a_1 + s_2 a_2)/(s_1 - s_2).$$

If we put $y = \{(x + d_1)/(x + d_2)\}^2$, then

$$R(x) = R_1(y) + R_2(y)y^{1/2},$$

where again the rational functions R_1, R_2 are determined by the rational function R, and

$$dx/g(x)^{1/2} = \pm dy/2|d_2 - d_1|[y(A_1y + B_1)(A_2y + B_2)]^{1/2}.$$

Thus we are again reduced to the case of a cubic with 3 distinct real roots.

The preceding argument may be applied also when g has only real roots, provided the factors Q_1 and Q_2 are chosen so that their zeros do not interlace. Suppose (without loss of generality) that $g = g_\lambda$ is in Riemann's normal form and take

$$Q_1 = (1 - x)(1 - \lambda x), \quad Q_2 = 4x.$$

In this case we can write

$$Q_1 = \{(1 + \sqrt{\lambda})^2(x - 1/\sqrt{\lambda})^2 - (1 - \sqrt{\lambda})^2(x + 1/\sqrt{\lambda})^2\}\sqrt{\lambda}/4,$$
$$Q_2 = -\sqrt{\lambda}\{(x - 1/\sqrt{\lambda})^2 - (x + 1/\sqrt{\lambda})^2\}.$$

If we put

$$1 - 4\sqrt{\lambda}y/(1 + \sqrt{\lambda})^2 = \{(x - 1/\sqrt{\lambda})/(x + 1/\sqrt{\lambda})\}^2,$$

we obtain

$$dx/g_\lambda(x)^{1/2} = dy/\mu g_\rho(y)^{1/2},$$

where

$$\mu = 1 + \sqrt{\lambda}, \quad \rho = 4\sqrt{\lambda}/(1 + \sqrt{\lambda})^2.$$

The usefulness of this change of variables will be seen in the next section.

2 The Arithmetic-Geometric Mean

Let a and b be positive real numbers, with $a > b$, and let

$$a_1 = (a + b)/2, \quad b_1 = (ab)^{1/2}$$

be respectively their arithmetic and geometric means. Then

$$a_1 < (a + a)/2 = a, \quad b_1 > (bb)^{1/2} = b,$$

and

$$a_1 - b_1 = (a^{1/2} - b^{1/2})^2/2 > 0.$$

Thus a_1, b_1 satisfy the same hypotheses as a, b and the procedure can be repeated. If we define sequences $\{a_n\}$, $\{b_n\}$ inductively by

$$a_0 = a, \quad b_0 = b,$$
$$a_{n+1} = (a_n + b_n)/2, \quad b_{n+1} = (a_n b_n)^{1/2} \quad (n = 0, 1, \ldots),$$

then

$$0 < b_0 < b_1 < b_2 < \cdots < a_2 < a_1 < a_0.$$

It follows that $a_n \to \lambda$ and $b_n \to \mu$ as $n \to \infty$, where $\lambda \geq \mu > 0$. In fact $\lambda = \mu$, as one sees by letting $n \to \infty$ in the relation $a_{n+1} = (a_n + b_n)/2$. The convergence of the sequences $\{a_n\}$ and $\{b_n\}$ to their common limit is extremely rapid, since

$$a_n - b_n = (a_{n-1} - b_{n-1})^2/8a_{n+1}.$$

(As an example, if $a = \sqrt{2}$ and $b = 1$, calculation shows that a_4 and b_4 differ by only one unit in the 20th decimal place.)

The common limit of the sequences $\{a_n\}$ and $\{b_n\}$ will be denoted by $M(a, b)$. The definition can be extended to arbitrary positive real numbers a, b by putting

$$M(a, a) = a, \quad M(b, a) = M(a, b).$$

Following Gauss (1818), $M(a, b)$ is known as the *arithmetic-geometric mean* of a and b. However, the preceding algorithm, which we will call the *AGM* algorithm, was first introduced by Lagrange (1784/5), who showed that it had a remarkable application to the numerical calculation of arbitrary elliptic integrals. The first tables of elliptic integrals, which made them as accessible as logarithms, were constructed in this way under the supervision of Legendre (1826). Today the algorithm can be used directly by electronic computers.

By putting $1 - \lambda x = t^2/a^2$ in Riemann's normal form, it may be seen that any real elliptic integral may be brought to the form

$$\int \varphi(t)[(a^2 - t^2)(t^2 - b^2)]^{-1/2} dt,$$

where $\varphi(t)$ is a rational function of t^2 with real coefficients, $a > b > 0$ and $t \in [b, a]$. We will restrict attention here to the *complete* elliptic integral

$$J = \int_b^a \varphi(t)[(a^2 - t^2)(t^2 - b^2)]^{-1/2} dt,$$

but at the cost of some complication the discussion may be extended to *incomplete* elliptic integrals (where the interval of integration is a proper subinterval of $[b, a]$).

If we make the change of variables

$$t^2 = a^2 \sin^2\theta + b^2 \cos^2\theta \quad (0 \leq \theta \leq \pi/2),$$

then

$$t\,dt/d\theta = (a^2 - b^2)\sin\theta\cos\theta = [(a^2 - t^2)(t^2 - b^2)]^{1/2}$$

and

$$J = \int_0^{\pi/2} \varphi((a^2\sin^2\theta + b^2\cos^2\theta)^{1/2})d\theta/(a^2\sin^2\theta + b^2\cos^2\theta)^{1/2}.$$

Now put

$$t_1 = (1/2)(t + ab/t)$$

and, as before,

$$a_1 = (a + b)/2, \quad b_1 = (ab)^{1/2}.$$

Then

$$\begin{aligned} a_1^2 - t_1^2 &= (a^2 - t^2)(t^2 - b^2)/4t^2, \\ t_1^2 - b_1^2 &= (t^2 - ab)^2/4t^2, \\ dt_1/dt &= (t^2 - ab)/2t^2. \end{aligned}$$

As t increases from b to b_1, t_1 decreases from a_1 to b_1, and as t further increases from b_1 to a, t_1 increases from b_1 back to a_1. Since

$$t = t_1 \pm (t_1^2 - b_1^2)^{1/2},$$

it follows from these observations that

$$\int_b^a \varphi(t)[(a^2 - t^2)(t^2 - b^2)]^{-1/2}dt = \int_{b_1}^{a_1} \psi(t_1)[(a_1^2 - t_1^2)(t_1^2 - b_1^2)]^{-1/2}dt_1,$$

where

$$\psi(t_1) = (1/2)\{\varphi[(t_1 + (t_1^2 - b_1^2)^{1/2}] + \varphi[(t_1 - (t_1^2 - b_1^2)^{1/2}]\}.$$

In particular, if we take $\varphi(t) = 1$ and put

$$\mathscr{K}(a, b) := \int_b^a [(a^2 - t^2)(t^2 - b^2)]^{-1/2}dt,$$

we obtain

$$\mathscr{K}(a, b) = \mathscr{K}(a_1, b_1).$$

Hence, by repeating the process, $\mathscr{K}(a, b) = \mathscr{K}(a_n, b_n)$. But

$$\mathscr{K}(a_n, b_n) = \int_0^{\pi/2} (a_n^2\sin^2\theta + b_n^2\cos^2\theta)^{-1/2}d\theta$$

and

$$b_n \le (a_n^2 \sin^2\theta + b_n^2 \cos^2\theta)^{1/2} \le a_n.$$

Consequently, by letting $n \to \infty$ we obtain

$$\mathscr{K}(a,b) = \pi/2M(a,b). \tag{1}$$

Now take $\varphi(t) = a^2 - t^2$ and put

$$\mathscr{E}(a,b) := \int_b^a [(a^2 - t^2)/(t^2 - b^2)]^{1/2} dt.$$

In this case

$$\psi(t_1) = (a^2 - b^2)/2 + 2(a_1^2 - t_1^2)$$

and hence

$$\mathscr{E}(a,b) = (a^2 - b^2)\mathscr{K}(a,b)/2 + 2\mathscr{E}(a_1,b_1).$$

If we write

$$e_n = 2^n(a_n^2 - b_n^2)$$

then, since $\mathscr{K}(a,b) = \mathscr{K}(a_n,b_n)$, by repeating the process we obtain

$$\mathscr{E}(a,b)/\mathscr{K}(a,b) = (e_0 + e_1 + \cdots + e_{n-1})/2 + 2^n\mathscr{E}(a_n,b_n)/\mathscr{K}(a_n,b_n).$$

But

$$2^n\mathscr{E}(a_n,b_n) = e_n \int_0^{\pi/2} \cos^2\theta (a_n^2 \sin^2\theta + b_n^2 \cos^2\theta)^{-1/2} d\theta$$

and $e_n \to 0$ (rapidly) as $n \to \infty$, since

$$e_n = 2^n(a_{n-1} - b_{n-1})^2/4 = e_{n-1}(a_{n-1} - b_{n-1})/4a_n.$$

Hence

$$\mathscr{E}(a,b)/\mathscr{K}(a,b) = (e_0 + e_1 + e_2 + \cdots)/2. \tag{2}$$

To avoid taking differences of nearly equal quantities, the constants e_n may be calculated by means of the recurrence relations

$$e_n = e_{n-1}^2/2^{n+2}a_n^2 \quad (n = 1, 2, \ldots).$$

Next take

$$\varphi(t) = p[(p^2 - a^2)(p^2 - b^2)]^{1/2}/(p^2 - t^2),$$

where either $p > a$ or $0 < p < b$, and put

$$\mathscr{P}(a,b,p) := \int_b^a p[(p^2 - a^2)(p^2 - b^2)]^{1/2} dt/(p^2 - t^2)[(a^2 - t^2)(t^2 - b^2)]^{1/2}.$$

In this case

$$\psi(t_1) = q_1 \pm p_1[(p_1^2 - a_1^2)(p_1^2 - b_1^2)]^{1/2}/(p_1^2 - t_1^2),$$

where

$$p_1 = (1/2)(p + ab/p),$$
$$q_1 = (p_1^2 - a_1^2)^{1/2} = [(p^2 - a^2)(p^2 - b^2)]^{1/2}/2p,$$

and the + or − sign is taken according as $p > a$ or $0 < p < b$. Since $p_1 > a_1$ in either event, without loss of generality *we now assume that* $p > a$. Then also $p_1 < p$ and

$$\mathscr{P}(a, b, p) = q_1 \mathscr{K}(a, b) + \mathscr{P}(a_1, b_1, p_1).$$

Define the sequence $\{p_n\}$ inductively by

$$p_0 = p, \quad p_{n+1} = (1/2)(p_n + a_n b_n / p_n) \ \ (n = 0, 1, \ldots),$$

and put

$$q_{n+1} = (p_{n+1}^2 - a_{n+1}^2)^{1/2} = [(p_n^2 - a_n^2)(p_n^2 - b_n^2)]^{1/2}/2p_n.$$

Then $p_n \to \upsilon \geq M(a, b)$ as $n \to \infty$, since $a_n < p_n < p_{n-1}$. In fact $\upsilon = M(a, b)$, as one sees by letting $n \to \infty$ in the recurrence relation defining the sequence $\{p_n\}$. Moreover

$$\delta_n := (a_n^2 - b_n^2)/(p_n^2 - a_n^2) \to 0 \text{ as } n \to \infty,$$

since

$$\delta_{n+1} = \delta_n \left(\frac{p_n^2}{4a_{n+1}^2}\right)\left(\frac{a_n^2 - b_n^2}{p_n^2 - b_n^2}\right) < \delta_n p_n^2/4a_{n+1}^2.$$

Hence

$$(p_n^2 - b_n^2)/(p_n^2 - a_n^2) = 1 + \delta_n \to 1.$$

Since $\mathscr{P}(a_n, b_n, p_n)$

$$= p_n[(p_n^2 - a_n^2)(p_n^2 - b_n^2)]^{1/2} \int_0^{\pi/2} \frac{(a_n^2 \sin^2\theta + b_n^2 \cos^2\theta)^{-1/2} d\theta}{(p_n^2 - a_n^2)\sin^2\theta + (p_n^2 - b_n^2)\cos^2\theta},$$

it follows that $\mathscr{P}(a_n, b_n, p_n) \to \pi/2$ as $n \to \infty$. Hence

$$\mathscr{P}(a, b, p) = (q_1 + q_2 + \cdots)\mathscr{K}(a, b) + \pi/2. \tag{3}$$

To avoid taking differences of nearly equal quantities, the constants q_n may be calculated by means of the recurrence relations

$$\delta_{n+1} = \delta_n^2 p_n^2/4a_{n+1}^2(1 + \delta_n), \quad q_{n+1} = (1 + \delta_n)^{1/2} q_n^2/2p_n \ \ (n = 1, 2, \ldots).$$

Using (1)–(3), complete elliptic integrals of all three kinds can be calculated by the *AGM* algorithm. We now consider another application, the utility of which will be seen in §6.

By putting $t_1 = (1/2)(t + ab/t)$ again, one sees that

$$\int_a^\infty [(t^2 - a^2)(t^2 - b^2)]^{-1/2} dt = (1/2) \int_{a_1}^\infty [(t_1^2 - a_1^2)(t_1^2 - b_1^2)]^{-1/2} dt_1.$$

But the change of variables $u = a(1 - b^2/t^2)^{1/2}$ shows that

$$\int_a^\infty [(t^2 - a^2)(t^2 - b^2)]^{-1/2} dt = \mathscr{K}(a, c),$$

where $c = (a^2 - b^2)^{1/2}$. It follows that

$$\mathscr{K}(a, c) = \mathscr{K}(a_1, c_1)/2 = \cdots = \mathscr{K}(a_n, c_n)/2^n,$$

where $c_n = (a_n^2 - b_n^2)^{1/2}$. The asymptotic behaviour of $\mathscr{K}(a_n, c_n)$ may be determined in the following way.

If we put $s = ac/t$, then s decreases from a to c as t increases from c to a, and

$$ds/dt = -[(a^2 - s^2)(s^2 - c^2)]^{1/2}/[(a^2 - t^2)(t^2 - c^2)]^{1/2}.$$

Since $s = t$ when $t = h := (ac)^{1/2}$, it follows that

$$\mathscr{K}(a, c) = 2 \int_c^h [(a^2 - t^2)(t^2 - c^2)]^{-1/2} dt.$$

But, for $c \le t \le h$,

$$b^{-1} = (a^2 - c^2)^{-1/2} \le (a^2 - t^2)^{-1/2} \le (a^2 - h^2)^{-1/2} = a^{-1}(1 - c/a)^{-1/2}.$$

Hence

$$2b^{-1}L \le \mathscr{K}(a, c) \le 2a^{-1}(1 - c/a)^{-1/2} L,$$

where

$$L := \int_c^h (t^2 - c^2)^{-1/2} dt = \log\{(a/c)^{1/2} + (a/c - 1)^{1/2}\}.$$

If we now replace a, b, c by a_n, b_n, c_n then, since $a_n/c_n \to \infty$ and moreover $a_n, b_n \to M(a, b)$, we deduce that

$$2^n \mathscr{K}(a, c)/\log(4a_n/c_n) \to 1/M(a, b) = 2\mathscr{K}(a, b)/\pi.$$

But $4a_n/c_n = (4a_n/c_{n-1})^2$, since $c_n = (a_{n-1} - b_{n-1})/2$, and hence

$$\begin{aligned} &2^{-n} \log(4a_n/c_n) \\ &\quad = 2^{1-n} \log(4a_{n-1}/c_{n-1}) - 2^{1-n} \log(a_{n-1}/a_n) \\ &\quad = \cdots \\ &\quad = \log(4a_0/c_0) - \log(a_0/a_1) - 2^{-1} \log(a_1/a_2) - \cdots - 2^{1-n} \log(a_{n-1}/a_n). \end{aligned}$$

It follows that

$$\pi \mathscr{K}(a,c)/2\mathscr{K}(a,b) = \log(4a_1/c_0) - \sum_{n=1}^{\infty} 2^{-n}\log(a_n/a_{n+1}). \tag{4}$$

Finally, to determine $\mathscr{E}(a,c)$ we can use the relation

$$\mathscr{K}(a,b)\mathscr{E}(a,c) + \mathscr{K}(a,c)\mathscr{E}(a,b) - a^2\mathscr{K}(a,b)\mathscr{K}(a,c) = \pi/2.$$

By homogeneity we need only establish this relation for $a = 1$. Since

$$\mathscr{K}(1,(1-\lambda)^{1/2}) = \int_0^1 [4x(1-x)(1-\lambda x)]^{-1/2}dx,$$
$$\mathscr{E}(1,(1-\lambda)^{1/2}) = \int_0^1 [(1-\lambda x)/4x(1-x)]^{1/2}dx,$$

it is in fact equivalent to the following relation, due to Legendre, between the complete elliptic integrals of the first and second kinds:

Proposition 1 *If*

$$K(\lambda) = \int_0^1 [4x(1-x)(1-\lambda x)]^{-1/2}dx, \quad E(\lambda) = \int_0^1 [(1-\lambda x)/4x(1-x)]^{1/2}dx,$$

then

$$K(\lambda)E(1-\lambda) + K(1-\lambda)E(\lambda) - K(\lambda)K(1-\lambda) = \pi/2 \quad \textit{for } 0 < \lambda < 1. \tag{5}$$

Proof We show first that the derivative of the left side of (5) is zero. Evidently

$$dE/d\lambda = -(1/2)\int_0^1 x[4x(1-x)(1-\lambda x)]^{-1/2}dx = [E(\lambda) - K(\lambda)]/2\lambda.$$

Similarly,

$$dK/d\lambda = (1/2)\int_0^1 x(1-\lambda x)^{-1}[4x(1-x)(1-\lambda x)]^{-1/2}dx.$$

Substituting $x = (1-u)/(1-\lambda u)$ and writing $\lambda' = 1-\lambda$, we obtain

$$\begin{aligned} dK/d\lambda &= (1/2\lambda')\int_0^1 [(1-u)/4u(1-\lambda u)]^{1/2}du \\ &= [E(\lambda) - \lambda' K(\lambda)]/2\lambda\lambda'. \end{aligned}$$

It follows that

$$d(\lambda\lambda' dK/d\lambda)/d\lambda = K/4.$$

Thus $y_1(\lambda) = K(\lambda)$ is a solution of the second order linear differential equation

$$d(\lambda\lambda' dy/d\lambda)/d\lambda - y/4 = 0. \tag{6}$$

By symmetry, $y_2(\lambda) = K(\lambda')$ is also a solution. It follows that the 'Wronskian'

$$W = \lambda\lambda'(y_2 dy_1/d\lambda - y_1 dy_2/d\lambda)$$

has derivative zero and so is constant. But, writing

$$K'(\lambda) = K(1-\lambda), \quad E'(\lambda) = E(1-\lambda),$$

we have

$$2W = K'(E - \lambda' K) + K(E' - \lambda K') = KE' + K'(E - K).$$

To evaluate this constant we let $\lambda \to 0$. Putting $x = \sin^2\theta$, we obtain

$$K(\lambda) = \int_0^{\pi/2} (1 - \lambda \sin^2\theta)^{-1/2} d\theta, \quad E(\lambda) = \int_0^{\pi/2} (1 - \lambda \sin^2\theta)^{1/2} d\theta$$

and hence, as $\lambda \to 0$,

$$K(\lambda) \to \pi/2, E(\lambda) \to \pi/2, E(\lambda') \to 1.$$

Moreover

$$K(\lambda')[E(\lambda) - K(\lambda)] \to 0,$$

since

$$K(\lambda) - E(\lambda) = \lambda \int_0^1 x[4x(1-x)(1-\lambda x)]^{-1/2} dx = O(\lambda)$$

and

$$0 \le K(\lambda') \le \int_0^{\pi/2} [1 - (1-\lambda)]^{-1/2} d\theta = O(\lambda^{-1/2}).$$

It follows that $2W = \pi/2$. □

If $\lambda = 1/2$, then $\lambda' = \lambda$ and (5) takes the simple form

$$K(1/2)[2E(1/2) - K(1/2)] = \pi/2.$$

By the remarks preceding the statement of Proposition 1, the left side can be evaluated by the *AGM* algorithm. In this way π has recently been calculated to millions of decimal places. (It will be recalled that the value $\lambda = 1/2$ occurred in the rectification of the lemniscate.)

3 Elliptic Functions

According to Jacobi, the theory of elliptic functions was conceived on 23 December 1751, the day on which the Berlin Academy asked Euler to report on the *Produzioni*

Matematiche of Count Fagnano, a copy of which had been sent them by the author. The papers which aroused Euler's interest had in fact already appeared in an obscure Italian journal between 1715 and 1720. Fagnano had shown first how a quadrant of a lemniscate could be halved, then how it could be divided algebraically into 2^m, $3 \cdot 2^m$ or $5 \cdot 2^m$ equal parts. He had also established an algebraic relation between the length of an elliptic arc, the length of another suitably chosen arc and the length of a quadrant. By analysing and extending his arguments, Euler was led ultimately (1761) to a general addition theorem for elliptic integrals. An elegant proof of Euler's theorem was given by Lagrange (1768/9), using differential equations. We follow this approach here.

Let

$$g_\lambda(x) = 4x(1-x)(1-\lambda x) = 4\lambda x^3 - 4(1+\lambda)x^2 + 4x$$

be Riemann's normal form and let $2f_\lambda(x)$ be its derivative:

$$f_\lambda(x) = 6\lambda x^2 - 4(1+\lambda)x + 2.$$

By the fundamental existence and uniqueness theorem for ordinary differential equations, the second order differential equation

$$x'' = f_\lambda(x) \tag{7}$$

has a unique solution $S(t) = S(t, \lambda)$, defined (and holomorphic) for $|t|$ sufficiently small, which satisfies the initial conditions

$$S(0) = S'(0) = 0. \tag{8}$$

The solution $S(t, \lambda)$ is an elementary function if $\lambda = 0$ or 1:

$$S(t, 0) = \sin^2 t, \quad S(t, 1) = \tanh^2 t.$$

(For other values of λ, $S(t)$ coincides with the Jacobian elliptic function $\mathrm{sn}^2 t$.)

Evidently $S(t)$ is an even function of t, since $S(-t)$ is also a solution of (7) and satisfies the same initial conditions (8).

For any solution $x(t)$ of (7), the function $x'(t)^2 - g_\lambda[x(t)]$ is a constant, since its derivative is zero. In particular,

$$S'(t)^2 = g_\lambda[S(t)], \tag{9}$$

since both sides vanish for $t = 0$.

If $|\tau|$ is sufficiently small, then $x_1(t) = S(t+\tau)$ and $x_2(t) = S(t-\tau)$ are solutions of (7) near $t = 0$. Moreover,

$$x_j'(t)^2 = g_\lambda[x_j(t)] \quad (j = 1, 2),$$

since these relations hold for $t = 0$. From

$$(x_1 x_2' + x_1' x_2)' = x_1 f_\lambda(x_2) + x_2 f_\lambda(x_1) + 2x_1' x_2'$$

and

$$(x_1 x_2' + x_1' x_2)^2 = x_1^2 g_\lambda(x_2) + x_2^2 g_\lambda(x_1) + 2x_1 x_2 x_1' x_2'$$

we obtain

$$2x_1x_2(x_1x_2' + x_1'x_2)' - (x_1x_2' + x_1'x_2)^2 - 2x_1x_2x_1'x_2'$$
$$= x_1^2\{2x_2 f_\lambda(x_2) - g_\lambda(x_2)\} + x_2^2\{2x_1 f_\lambda(x_1) - g_\lambda(x_1)\}.$$

But if $g_\lambda(x) = \alpha x^3 + \beta x^2 + \gamma x$ and $f_\lambda(x) = g_\lambda'(x)/2$, then

$$2x f_\lambda(x) - g_\lambda(x) = x^2(2\alpha x + \beta).$$

Hence

$$2x_1x_2(x_1x_2' + x_1'x_2)' - (x_1x_2' + x_1'x_2)^2 = 2x_1^2x_2^2\{\alpha(x_1 + x_2) + \beta\} + 2x_1x_2x_1'x_2'.$$

On the other hand,

$$(x_1' - x_2')(x_1x_2' + x_1'x_2) = x_2g_\lambda(x_1) - x_1g_\lambda(x_2) + (x_1 - x_2)x_1'x_2'$$
$$= x_1x_2(x_1 - x_2)\{\alpha(x_1 + x_2) + \beta\} + (x_1 - x_2)x_1'x_2'.$$

Comparing these two relations, we obtain

$$\{2x_1x_2(x_1x_2' + x_1'x_2)' - (x_1x_2' + x_1'x_2)^2\}(x_1 - x_2) = 2x_1x_2(x_1' - x_2')(x_1x_2' + x_1'x_2).$$

If we divide by $2x_1x_2(x_1 - x_2)(x_1x_2' + x_1'x_2)$, this takes the form

$$\frac{(x_1x_2' + x_1'x_2)'}{x_1x_2' + x_1'x_2} - \frac{x_1x_2' + x_1'x_2}{2x_1x_2} = \frac{x_1' - x_2'}{x_1 - x_2},$$

which can be integrated to give

$$(x_1x_2' + x_1'x_2)^2 = C(\tau)x_1x_2(x_1 - x_2)^2,$$

where the constant $C(\tau)$ depends on τ. Equivalently,

$$[S(u)S'(v) - S'(u)S(v)]^2 = C((u + v)/2)S(u)S(v)[S(u) - S(v)]^2.$$

To evaluate the constant, we divide throughout by $S(v)$ and let $v \to 0$. By (9), this yields $C(u/2) = \gamma/S(u)$. Since $\gamma = 4$ (for Riemann's normal form), we obtain finally

$$S(u + v) = 4S(u)S(v)[S(u) - S(v)]^2/[S(u)S'(v) - S'(u)S(v)]^2. \tag{10}$$

Thus $S(u + v)$ is a rational function of $S(u)$, $S(v)$, $S'(u)$, $S'(v)$. Moreover, since $(S')^2 = g_\lambda(S)$, there exists a polynomial $p(x, y, z)$, not identically zero and with coefficients independent of u and v, such that $p[S(u+v), S(u), S(v)] = 0$. In other words, the function $S(u)$ has an *algebraic addition theorem*.

The relation (10) can also be written in the form

$$S(u + v) = [S(u)S'(v) + S'(u)S(v)]^2/4S(u)S(v)[1 - \lambda S(u)S(v)]^2, \tag{11}$$

since

$$S(u)^2S'(v)^2 - S'(u)^2S(v)^2 = S(u)^2 g_\lambda[S(v)] - S(v)^2 g_\lambda[S(u)]$$
$$= 4S(u)S(v)[S(u) - S(v)][1 - \lambda S(u)S(v)].$$

Replacing v by $-v$ in (11) and subtracting the result from (11), we obtain

$$S(u+v) - S(u-v) = S'(u)S'(v)/[1 - \lambda S(u)S(v)]^2. \tag{12}$$

In particular, for $v = u$,

$$S(2u) = g_\lambda[S(u)]/[1 - \lambda S^2(u)]^2. \tag{13}$$

We recall that a function is *meromorphic* in a connected open set D if it is holomorphic throughout D, except for isolated singularities which are poles. Since, by (13), $S(2t)$ is a rational function of $S(t)$, it follows that if $S(t)$ is meromorphic and a solution (wherever it is finite) of the differential equation (7) in an open disc $|t| < R$, then its definition can be extended so that it is meromorphic and a solution (wherever it is finite) of the differential equation (7) also in the disc $|t| < 2R$. But the fundamental existence and uniqueness theorem guarantees that $S(t)$ is holomorphic in a neighbourhood of the origin. Consequently we can extend its definition so that it is meromorphic and a solution of (7) in the whole complex plane $\mathbb{C}$.

Further properties of the function $S(t)$ may be derived from the differential equation (7). For any constants α, β, if $y(t) = \alpha S(\beta t)$, then $y(0) = y'(0) = 0$. It is readily seen that $y(t)$ satisfies a differential equation of the form (7) if and only if either $\alpha = 1$, $\beta = \pm 1$ or $\alpha = \lambda$, $\lambda\beta^2 = 1$, and in the latter case with λ replaced by $1/\lambda$ in (7). It follows that, for any $\lambda \neq 0$,

$$S(t, 1/\lambda) = \lambda S(\lambda^{-1/2}t, \lambda). \tag{14}$$

By differentiation it may be shown also that $S(it, \lambda)/[S(it, \lambda) - 1]$, where $i^2 = -1$, is a solution of the differential equation (7) with λ replaced by $1 - \lambda$. It follows that

$$S(t, 1-\lambda) = S(it, \lambda)/[S(it, \lambda) - 1]. \tag{15}$$

By combining (14) and (15) we obtain, for any $\lambda \neq 0, 1$, three more relations:

$$S(t, 1/(1-\lambda)) = (1-\lambda)S(i(1-\lambda)^{-1/2}t, \lambda)/[S(i(1-\lambda)^{-1/2}t, \lambda) - 1], \tag{16}$$

$$S(t, (\lambda-1)/\lambda) = \lambda S(i\lambda^{-1/2}t, \lambda)/[\lambda S(i\lambda^{-1/2}t, \lambda) - 1], \tag{17}$$

$$S(t, \lambda/(\lambda-1)) = (1-\lambda)S((1-\lambda)^{-1/2}t, \lambda)/[1 - \lambda S((1-\lambda)^{-1/2}t, \lambda)]. \tag{18}$$

As in §1, it follows from (14)–(18) that the evaluation of $S(t, \lambda)$ for all $t, \lambda \in \mathbb{C}$ reduces to its evaluation for λ in the region $|\lambda - 1| \leq 1$, $0 \leq \mathscr{R}\lambda \leq 1/2$. Similarly it follows from (14) and (18) that the evaluation of $S(t, \lambda)$ for all $t, \lambda \in \mathbb{R}$ reduces to its evaluation for λ in the interval $0 < \lambda < 1$. We now show that $S(t, \lambda)$ can then be calculated by the *AGM* algorithm.

It is easily verified that if

$$z(t) = (1 + \sqrt{\lambda})^2 S(t, \lambda)/[1 + \sqrt{\lambda}S(t, \lambda)]^2,$$

then

$$(dz/dt)^2 = (1+\sqrt{\lambda})^2\{4\lambda_0 z^3 - 4(1+\lambda_0)z^2 + 4z\},$$

where

$$\lambda_0 = 4\sqrt{\lambda}/(1+\sqrt{\lambda})^2. \tag{19}$$

Since $z(0) = z'(0) = 0$ and $z''(0) \neq 0$, it follows that $z(t) = S((1+\sqrt{\lambda})t, \lambda_0)$. Thus

$$S((1+\sqrt{\lambda})t, \lambda_0) = (1+\sqrt{\lambda})^2 S(t,\lambda)/[1+\sqrt{\lambda}S(t,\lambda)]^2. \tag{20}$$

The inequality $0 < \lambda < 1$ implies $\lambda < \lambda_0 < 1$. Hence, by regarding (19) as a quadratic equation for $\sqrt{\lambda}$, we obtain

$$\sqrt{\lambda} = [1-(1-\lambda_0)^{1/2}]^2/\lambda_0. \tag{21}$$

If we write $\sqrt{\lambda_0} = c_0/a_0$, where $c_0 = (a_0^2 - b_0^2)^{1/2}$ and $0 < b_0 < a_0$, then

$$\sqrt{\lambda} = (a_0 - b_0)/(a_0 + b_0) = c_1/a_1,$$

where

$$a_1 = (a_0+b_0)/2, \quad b_1 = (a_0 b_0)^{1/2}, \quad c_1 = (a_1^2 - b_1^2)^{1/2}.$$

Since $1+\sqrt{\lambda} = a_0/a_1$, we can rewrite (20) in the form

$$S(a_0 t, \lambda_0) = (1+c_1/a_1)^2 S(a_1 t, \lambda_1)/[1+(c_1/a_1)S(a_1 t, \lambda_1)]^2,$$

where $\lambda_1 = \lambda = (c_1/a_1)^2$. Repeating the process, we obtain

$$S(a_{n-1}t, \lambda_{n-1}) = (1+c_n/a_n)^2 S(a_n t, \lambda_n)/[1+(c_n/a_n)S(a_n t, \lambda_n)]^2,$$

where $\lambda_n = (c_n/a_n)^2$. As $n \to \infty$,

$$a_n \to \mu := M(a,b), \quad c_n \to 0, \quad \lambda_n \to 0.$$

Since $S(t,0) = \sin^2 t$, for some (not very large) $n = N$ we have $S(a_N t, \lambda_N) \approx \sin^2 \mu t$, which may be considered as known. Then, by taking successively $n = N, N-1, \ldots, 1$ we can calculate $S(a_0 t, \lambda_0)$. Moreover, we can start the process by taking $a_0 = 1$, $b_0 = (1-\lambda_0)^{1/2}$.

We now consider periodicity properties. If $\lambda \neq 1$ and $S(h) = 1$ for some nonzero $h \in \mathbb{C}$ then, by (13), $S(2h) = 0$. Furthermore $S'(2h) = 0$, by (9). It follows that $S(t)$ has period $2h$, since $S(t+2h)$ is a solution of the differential equation (7) which satisfies the same initial conditions (8) as $S(t)$. It remains to show that there exists such an h.

Suppose first that $\lambda \in \mathbb{R}$ and $0 < \lambda < 1$. Since $S''(0) = 2$, we have $S'(t) > 0$ for small $t > 0$. If $S'(t) > 0$ for $0 < t < T$, then $S(t)$ is a positive increasing function for $0 < t < T$. Since $g_\lambda[S(t)] > 0$, we must also have $S(t) < 1$ for $0 < t < T$. From the relation

$$t = \int_0^{S(t)} dx/g_\lambda(x)^{1/2},$$

it follows that $T \leq K(\lambda)$, where

$$K(\lambda) := \int_0^1 dx/g_\lambda(x)^{1/2}.$$

Hence $S'(t)$ vanishes for some t such that $0 < t \leq K(\lambda)$ and we can now take T to be the least $t > 0$ for which $S'(t) = 0$. Then $S'(T) = 0$, $S(T) = 1$ and by letting $t \to T$ we obtain $T = K(\lambda)$.

This shows that $S(u)$ maps the interval $[0, K(\lambda)]$ bijectively onto $[0, 1]$, and if

$$u(\xi) = \int_0^\xi dx/g_\lambda(x)^{1/2} \quad (0 \leq \xi \leq 1),$$

then $S[u(\xi)] = \xi$. Thus, in the real domain, the elliptic integral of the first kind is *inverted* by the function $S(u)$.

Since $\lambda \neq 1$, it follows that $S(t) = S(t, \lambda)$ has period $2K(\lambda)$. Since $\lambda \neq 0$, it follows from (15) that $S(t, \lambda)$ also has period $2iK(1 - \lambda)$. Thus $S(t, \lambda)$ is a *doubly-periodic* function, with a real period and a pure imaginary period. We will show that all periods are given by

$$2mK(\lambda) + 2niK(1 - \lambda) \quad (m, n \in \mathbb{Z}).$$

The periods of a nonconstant meromorphic function f form a discrete additive subgroup of $\mathbb{C}$. If f has two periods whose ratio is not real then, by the simple case $n = 2$ of Proposition VIII.7, it has periods ω_1, ω_2 such that all periods are given by

$$m\omega_1 + n\omega_2 \quad (m, n \in \mathbb{Z}).$$

In the present case we can take $\omega_1 = 2K(\lambda)$, $\omega_2 = 2iK(1 - \lambda)$ since, by construction, $2K(\lambda)$ is the least positive period.

Suppose next that $\lambda \in \mathbb{R}$ and either $\lambda > 1$ or $\lambda < 0$. Then, by (14) and (15), $S(t, \lambda)$ is again a doubly-periodic function with a real period and a pure imaginary period.

Suppose finally that $\lambda \in \mathbb{C}\backslash\mathbb{R}$. Without loss of generality, *we assume* $\mathscr{I}\lambda > 0$. Then $g_\lambda(z)$ does not vanish in the upper half-plane $\mathscr{H}$. It follows that there exists a unique function $h_\lambda(z)$, holomorphic for $z \in \mathscr{H}$ with $\mathscr{R}h_\lambda(z) > 0$ for z near 0, such that

$$h_\lambda(z)^2 = g_\lambda(z). \tag{22}$$

Moreover, we may extend the definition so that $h_\lambda(z)$ is continuous and (22) continues to hold for $z \in \mathscr{H} \cup \mathbb{R}$.

We can write $S(t) = \psi(t^2)$, where

$$\psi(w) = w + a_2 w^2 + \cdots$$

is holomorphic at the origin. By inversion of series, there exists a function

$$\phi(z) = z + b_2 z^2 + \cdots,$$

which is holomorphic at the origin, such that $\psi[\phi(z)] = z$. For $z \in \mathscr{H}$ near 0, put

$$u(z) = \phi(z)^{1/2},$$

where the square root is chosen so that $\mathscr{R}u(z) > 0$. Then $S[u(z)] = z$. Differentiating and then squaring, we obtain

$$S'[u(z)]u'(z) = 1, \quad u'(z)^2 = 1/g_\lambda(z).$$

But $u'(z)$ also has positive real part, since $S'[u(z)] \sim 2u(z)$ for $z \to 0$. Consequently $u'(z) = 1/h_\lambda(z)$. Since $u(z) \to 0$ as $z \to 0$, we conclude that

$$u(z) = \int_0^z d\zeta / h_\lambda(\zeta), \tag{23}$$

where the path of integration is (say) a straight line segment. However, the function on the right is holomorphic for all $z \in \mathscr{H}$. Consequently, if we define $u(z)$ by (23) then, by analytic continuation, the relation $S[u(z)] = z$ continues to hold for all $z \in \mathscr{H}$. Letting $z \to 1$, we now obtain $S(h) = 1$ for $h = K(\lambda)$, where

$$K(\lambda) := \int_0^1 dx / g_\lambda(x)^{1/2}$$

and the square root is chosen so that $g_\lambda(x)^{1/2}$ is continuous and has positive real part for small $x > 0$ and actually, as we will see in a moment, for $0 < x < 1$. Hence $S(t)$ has period $2K(\lambda)$. Furthermore, by (15), $S(t)$ also has period $2iK(1-\lambda)$.

For $0 < x < 1$ we have

$$1/g_\lambda(x)^{1/2} = (1-\bar{\lambda}x)^{1/2}/[4x(1-x)]^{1/2}|1-\lambda x|.$$

If $\lambda = \mu + iv$, where $v > 0$, then $1 - \bar{\lambda}x = \gamma + i\delta$, where $\gamma = 1 - \mu x$ and $\delta = vx > 0$ for $0 < x < 1$. Hence

$$(1-\bar{\lambda}x)^{1/2} = \alpha + i\beta,$$

where

$$\alpha = \{\gamma + (\gamma^2 + \delta^2)^{1/2}\}^{1/2}/\sqrt{2}, \quad 2\alpha\beta = \delta,$$

first for small $x > 0$ and then, by continuity, for $0 < x < 1$. Thus α and β are positive for $0 < x < 1$. Consequently $\mathscr{R}g_\lambda(x)^{1/2} > 0$ for $0 < x < 1$ and

$$K(\lambda) = A + iB,$$

where $A > 0$, $B > 0$.

Similarly, for $0 < y < 1$ we have

$$1/g_{1-\lambda}(y)^{1/2} = (1-(1-\bar{\lambda})y)^{1/2}/[4y(1-y)]^{1/2}|1-(1-\lambda)y|$$

and $1-(1-\bar{\lambda})y = \gamma' - i\delta'$, where $\gamma' = 1-(1-\mu)y$ and $\delta' = vy > 0$ for $0 < y < 1$. Hence

$$(1-(1-\bar{\lambda})y)^{1/2} = \alpha' - i\beta',$$

where

$$\alpha' = \{\gamma' + (\gamma'^2 + \delta'^2)^{1/2}\}^{1/2}/\sqrt{2}, \quad 2\alpha'\beta' = \delta'.$$

Thus α' and β' are positive for $0 < y < 1$, and

$$K(1-\lambda) = A' - iB',$$

where $A' > 0$, $B' > 0$.

We will now show that the period ratio $iK(1-\lambda)/K(\lambda)$ is not real by showing that the quotient $K(1-\lambda)/K(\lambda)$ has positive real part. Since this is equivalent to showing that

$$AA' - BB' > 0,$$

it is sufficient to show that $\alpha\alpha' - \beta\beta' > 0$ for all $x, y \in (0, 1)$. The inequality is certainly satisfied for all x, y near 0, since $\alpha \to 1$, $\beta \to 0$ as $x \to 0$ and $\alpha' \to 1$, $\beta' \to 0$ as $y \to 0$. Thus we need only show that we never have $\alpha\alpha' = \beta\beta'$. But

$$2\alpha^2 = (\gamma^2 + \delta^2)^{1/2} + \gamma, \quad 2\beta^2 = (\gamma^2 + \delta^2)^{1/2} - \gamma,$$

with analogous expressions for $2\alpha'^2, 2\beta'^2$. Hence, if $\alpha\alpha' = \beta\beta'$, then by squaring we obtain

$$[(\gamma^2 + \delta^2)^{1/2} + \gamma][(\gamma'^2 + \delta'^2)^{1/2} + \gamma'] = [(\gamma^2 + \delta^2)^{1/2} - \gamma][(\gamma'^2 + \delta'^2)^{1/2} - \gamma'],$$

which reduces to

$$\gamma(\gamma'^2 + \delta'^2)^{1/2} = -\gamma'(\gamma^2 + \delta^2)^{1/2}.$$

Squaring again, we obtain $\gamma^2\delta'^2 = \gamma'^2\delta^2$. Since the previous equation shows that γ and γ' do not have the same sign, it follows that

$$\gamma\delta' + \gamma'\delta = 0.$$

Giving $\gamma, \delta, \gamma', \delta'$ their explicit expressions, this takes the form $v(x + y - xy) = 0$. Hence $x(1-y) + y = 0$, which is impossible if $0 < y < 1$ and $x > 0$.

The relation $S[u(z)] = z$, where $u(z)$ is defined by (23), shows that the elliptic integral of the first kind is *inverted* by the elliptic function $S(u)$. We may use this to simplify other elliptic integrals. The change of variables $x = S(u)$ replaces the integral

$$\int R(x)\,dx/g_\lambda(x)^{1/2}$$

by $\int R[S(u)]du$. Following Jacobi, we take

$$E(u) := \int_0^u [1 - \lambda S(v)]dv \tag{24}$$

as the standard elliptic integral of the second kind, and

$$\Pi(u,a) := (\lambda/2)\int_0^u S'(a)S(v)dv/[1-\lambda S(a)S(v)] \tag{25}$$

as the standard elliptic integral of the third kind.

Many properties of these functions may be obtained by integration from corresponding properties of the function $S(u)$. By way of example, we show that

$$E(u+a)-E(u-a)-2E(a) = -\lambda S'(a)S(u)/[1-\lambda S(a)S(u)]. \tag{26}$$

Indeed it is evident that both sides vanish when $u=0$, and it follows from (12) that they have the same derivative with respect to u. Integrating (26) with respect to u, we further obtain

$$\Pi(u,a) = uE(a) - (1/2)\int_{u-a}^{u+a} E(v)dv. \tag{27}$$

Thus the function $\Pi(u,a)$, which depends on two variables (as well as the parameter λ) can be expressed in terms of functions of only one variable. Furthermore, we have the *interchange property* (due, in other notation, to Legendre)

$$\Pi(u,a) - uE(a) = \Pi(a,u) - aE(u). \tag{28}$$

If we take $u = 2K = 2K(\lambda)$, then $S'(u)=0$ and hence $\Pi(a,u)=0$. Thus

$$\Pi(2K,a) = 2KE(a) - aE(2K), \tag{29}$$

which shows that the complete elliptic integral of the third kind can be expressed in terms of complete and incomplete elliptic integrals of the first and second kinds.

In order to justify taking $\Pi(u,a)$ as the standard elliptic integral of the third kind, we show finally that $S(a)$ takes all complex values. Otherwise, if $S(u)\neq c$ for all $u\in\mathbb{C}$, then $c\neq 0$ and

$$f(u) = S(u)/[S(u)-c]$$

is holomorphic in the whole complex plane. Furthermore, it is doubly-periodic with two periods ω_1,ω_2 whose ratio is not real. Since it is bounded in the parallelogram with vertices $0,\omega_1,\omega_2,\omega_1+\omega_2$, it follows that it is bounded in $\mathbb{C}$. Hence, by Liouville's theorem, f is a constant. Since S is not constant and $c\neq 0$, this is a contradiction.

4 Theta Functions

Theta functions arise not only in connection with elliptic functions (as we will see), but also in problems of heat conduction, statistical mechanics and number theory.

Consider the bi-infinite series

$$\sum_{n=-\infty}^{\infty} q^{n^2}z^n = 1 + \sum_{n=1}^{\infty} q^{n^2}z^n + \sum_{n=1}^{\infty} q^{n^2}z^{-n},$$

where $q, z \in \mathbb{C}$ and $z \neq 0$. Both series on the right converge if $|q| < 1$, both diverge if $|q| > 1$, and at most one converges if $|q| = 1$. Thus we now assume $|q| < 1$.

A remarkable representation for the series on the left was given by Jacobi (1829), in §64 of his *Fundamenta Nova*, and is now generally known as *Jacobi's triple product formula*:

Proposition 2 *If $|q| < 1$ and $z \neq 0$, then*

$$\sum_{n=-\infty}^{\infty} q^{n^2} z^n = \prod_{n=1}^{\infty} (1 + q^{2n-1}z)(1 + q^{2n-1}z^{-1})(1 - q^{2n}). \tag{30}$$

Proof Put

$$f_N(z) = \prod_{n=1}^{N} (1 + q^{2n-1}z)(1 + q^{2n-1}z^{-1}).$$

Then we can write

$$f_N(z) = c_0^N + c_1^N (z + z^{-1}) + \cdots + c_N^N (z^N + z^{-N}). \tag{31}$$

To determine the coefficients c_n^N we use the functional relation

$$\begin{aligned} f_N(q^2 z) &= (1 + q^{2N+1}z)(1 + q^{-1}z^{-1}) f_N(z)/(1 + qz)(1 + q^{2N-1}z^{-1}) \\ &= (1 + q^{2N+1}z) f_N(z)/(qz + q^{2N}). \end{aligned}$$

Multiplying both sides by $qz + q^{2N}$ and equating coefficients of z^{n+1} we get, for $n = 0, 1, \ldots, N-1$,

$$q^{2n+1} c_n^N + q^{2N+2n+2} c_{n+1}^N = c_{n+1}^N + q^{2N+1} c_n^N,$$

i.e.,

$$q^{2n+1}(1 - q^{2N-2n}) c_n^N = (1 - q^{2N+2n+2}) c_{n+1}^N.$$

But, since $\sum_{n=1}^{N} (2n-1) = N^2$, it follows from the definition of $f_N(z)$ that $c_N^N = q^{N^2}$. Hence, for $0 \leq n \leq N$,

$$c_n^N = (1 - q^{2N+2n+2})(1 - q^{2N+2n+4}) \cdots (1 - q^{4N}) q^{n^2} / D,$$

where $D = (1 - q^2)(1 - q^4) \cdots (1 - q^{2N-2n})$.

If $|q| < 1$ and $z \neq 0$, then the infinite products

$$\prod_{n=1}^{\infty} (1 + q^{2n-1}z), \prod_{n=1}^{\infty} (1 + q^{2n-1}z^{-1}), \prod_{n=1}^{\infty} (1 - q^{2n})$$

are all convergent. From the convergence of the last it follows that, for each fixed n,

$$\lim_{N \to \infty} c_n^N = q^{n^2} \Big/ \prod_{k=1}^{\infty} (1 - q^{2k}).$$

Moreover, there exists a constant $A > 0$, depending on q but not on n or N, such that

$$|c_n^N| \leq A|q|^{n^2}.$$

For we can choose $B > 0$ so that $|\prod_{k=1}^{m}(1-q^{2k})| \geq B$ for all m, we can choose $C > 0$ so that $|\prod_{k=1}^{m}(1-q^{2k})| \leq C$ for all m, and we can then take $A = C/B^2$. Since the series $\sum_{n=-\infty}^{\infty} q^{n^2} z^n$ is absolutely convergent, it follows that we can proceed to the limit term by term in (31) to obtain (30). □

In the series $\sum_{n=-\infty}^{\infty} q^{n^2} z^n$ we now put

$$q = e^{\pi i \tau}, \quad z = e^{2\pi i v},$$

so that $|q| < 1$ corresponds to $\mathscr{I}\tau > 0$, and we define the *theta function*

$$\theta(v;\tau) = \sum_{n=-\infty}^{\infty} e^{\pi i \tau n^2} e^{2\pi i v n}.$$

The function $\theta(v;\tau)$ is holomorphic in v and τ for all $v \in \mathbb{C}$ and $\tau \in \mathscr{H}$ (the upper half-plane). Since initially we will be more interested in the dependence on v, with τ just a parameter, we will often write $\theta(v)$ in place of $\theta(v;\tau)$. Furthermore, we will still use q as an abbreviation for $e^{\pi i \tau}$.

Evidently

$$\theta(v+1) = \theta(v) = \theta(-v).$$

Moreover,

$$\begin{aligned}\theta(v+\tau) &= \sum_{n=-\infty}^{\infty} q^{n^2+2n} e^{2\pi i v n} \\ &= q^{-1} e^{-2\pi i v} \sum_{n=-\infty}^{\infty} q^{(n+1)^2} e^{2\pi i v (n+1)} \\ &= e^{-\pi i (2v+\tau)} \theta(v).\end{aligned}$$

It may be immediately verified that

$$\partial^2\theta/\partial v^2 = -4\pi^2 q \partial\theta/\partial q = 4\pi i \partial\theta/\partial\tau,$$

which becomes the partial differential equation of heat conduction in one dimension on putting $\tau = 4\pi i t$.

By Proposition 2, we have also the product representation

$$\theta(v) = \prod_{n=1}^{\infty}(1+q^{2n-1}e^{2\pi i v})(1+q^{2n-1}e^{-2\pi i v})(1-q^{2n}).$$

It follows that the points

$$v = 1/2 + \tau/2 + m + n\tau \quad (m, n \in \mathbb{Z})$$

are simple zeros of $\theta(v)$, and that these are the only zeros.

One important property of the theta function is almost already known to us:

Proposition 3 *For all $v \in \mathbb{C}$ and $\tau \in \mathscr{H}$,*

$$\theta(v; -1/\tau) = (\tau/i)^{1/2} e^{\pi i \tau v^2} \theta(\tau v; \tau), \tag{32}$$

where the square root is chosen to have positive real part.

Proof Suppose first that $\tau = iy$, where $y > 0$. We wish to show that

$$\sum_{n=-\infty}^{\infty} e^{-n^2\pi/y} e^{2n\pi i v} = y^{1/2} \sum_{n=-\infty}^{\infty} e^{-(v+n)^2 \pi y}.$$

But this was already proved in Proposition IX.10.

Thus (32) holds when τ is pure imaginary. Since, with the stated choice of square root, both sides of (32) are holomorphic functions for $v \in \mathbb{C}$ and $\tau \in \mathscr{H}$, the relation continues to hold throughout this extended domain, by analytic continuation. □

Following Hermite (1858), for any integers α, β we now put

$$\theta_{\alpha,\beta}(v) = \theta_{\alpha,\beta}(v; \tau) = \sum_{n=-\infty}^{\infty} (-1)^{\beta n} e^{\pi i \tau (n+\alpha/2)^2} e^{2\pi i v (n+\alpha/2)}.$$

(The factor $(-1)^{\beta n}$ may be made less conspicuous by writing it as $e^{\pi i \beta n}$.) Since

$$\theta_{\alpha+2,\beta}(v) = (-1)^{\beta} \theta_{\alpha,\beta}(v), \quad \theta_{\alpha,\beta+2}(v) = \theta_{\alpha,\beta}(v),$$

there are only four essentially distinct functions, namely

$$\begin{aligned}
\theta_{00}(v) &= \sum_{n=-\infty}^{\infty} e^{\pi i \tau n^2} e^{2\pi i v n}, \\
\theta_{01}(v) &= \sum_{n=-\infty}^{\infty} (-1)^n e^{\pi i \tau n^2} e^{2\pi i v n}, \\
\theta_{10}(v) &= \sum_{n=-\infty}^{\infty} e^{\pi i \tau (n+1/2)^2} e^{\pi i v (2n+1)}, \\
\theta_{11}(v) &= \sum_{n=-\infty}^{\infty} (-1)^n e^{\pi i \tau (n+1/2)^2} e^{\pi i v (2n+1)}.
\end{aligned} \tag{33}$$

Moreover,

$$\begin{gathered}
\theta_{00}(v; \tau) = \theta(v; \tau), \quad \theta_{01}(v; \tau) = \theta(v + 1/2; \tau), \\
\theta_{10}(v; \tau) = e^{\pi i (v+\tau/4)} \theta(v + \tau/2; \tau), \quad \theta_{11}(v; \tau) = e^{\pi i (v+\tau/4)} \theta(v + 1/2 + \tau/2; \tau).
\end{gathered}$$

In fact, for all integers m, n,

$$\theta_{\alpha,\beta}(v + m\tau/2 + n/2) = \theta_{\alpha+m,\beta+n}(v) e^{-\pi i (mv + m^2\tau/4 - \alpha n/2)}. \tag{34}$$

Since the zeros of $\theta(v;\tau)$ are the points $v = 1/2 + \tau/2 + m\tau + n$, the zeros of $\theta_{\alpha,\beta}(v)$ are the points

$$v = (\beta+1)/2 + (\alpha+1)\tau/2 + m\tau + n \ (m, n \in \mathbb{Z}).$$

The notation for theta functions is by no means standardized. Hermite's notation reflects the underlying symmetry, but for purposes of comparison we indicate its connection with the more commonly used notation in Whittaker and Watson [29]:

$$\theta_{00}(v;\tau) = \vartheta_3(\pi v, q), \quad \theta_{01}(v;\tau) = \vartheta_4(\pi v, q),$$
$$\theta_{10}(v;\tau) = \vartheta_2(\pi v, q), \quad \theta_{11}(v;\tau) = i\vartheta_1(\pi v, q).$$

It follows from the definitions that $\theta_{00}(v;\tau)$, $\theta_{01}(v;\tau)$ and $\theta_{10}(v;\tau)$ are even functions of v, whereas $\theta_{11}(v;\tau)$ is an odd function of v. Moreover $\theta_{00}(v;\tau)$ and $\theta_{01}(v;\tau)$ are periodic with period 1 in v, but $\theta_{10}(v;\tau)$ and $\theta_{11}(v;\tau)$ change sign when v is increased by 1.

All four theta functions satisfy the same partial differential equation as $\theta(v;\tau)$. From the product expansion of $\theta(v;\tau)$ we obtain the product expansions

$$\begin{aligned}
\theta_{00}(v) &= Q_0 \prod_{n=1}^{\infty}(1+q^{2n-1}e^{2\pi iv})(1+q^{2n-1}e^{-2\pi iv}),\\
\theta_{01}(v) &= Q_0 \prod_{n=1}^{\infty}(1-q^{2n-1}e^{2\pi iv})(1-q^{2n-1}e^{-2\pi iv}),\\
\theta_{10}(v) &= 2Q_0 e^{\pi i\tau/4}\cos\pi v \prod_{n=1}^{\infty}(1+q^{2n}e^{2\pi iv})(1+q^{2n}e^{-2\pi iv}),\\
\theta_{11}(v) &= 2iQ_0 e^{\pi i\tau/4}\sin\pi v \prod_{n=1}^{\infty}(1-q^{2n}e^{2\pi iv})(1-q^{2n}e^{-2\pi iv}),
\end{aligned} \tag{35}$$

where $q = e^{\pi i\tau}$ *and*

$$Q_0 = \prod_{n=1}^{\infty}(1-q^{2n}).$$

In particular,

$$\theta_{00}(0) = Q_0 \prod_{n=1}^{\infty}(1+q^{2n-1})^2,$$
$$\theta_{01}(0) = Q_0 \prod_{n=1}^{\infty}(1-q^{2n-1})^2,$$
$$\theta_{10}(0) = 2q^{1/4}Q_0 \prod_{n=1}^{\infty}(1+q^{2n})^2.$$

By differentiating with respect to v and then putting $v = 0$, we obtain in addition $\theta_{11}'(0) = 2\pi i q^{1/4} Q_0^3$. But

$$\begin{aligned} Q_0 &= \prod_{n=1}^{\infty} (1-q^n)(1+q^n) \\ &= \prod_{n=1}^{\infty} (1-q^{2n})(1-q^{2n-1})(1+q^{2n})(1+q^{2n-1}), \end{aligned}$$

which implies

$$\prod_{n=1}^{\infty} (1-q^{2n-1})(1+q^{2n})(1+q^{2n-1}) = 1.$$

It follows that

$$\theta_{00}(0)\theta_{01}(0)\theta_{10}(0) = 2q^{1/4} Q_0^3$$

and hence

$$\theta_{11}'(0) = \pi i \theta_{00}(0)\theta_{01}(0)\theta_{10}(0). \tag{36}$$

It is evident from their series definitions that, when q is replaced by $-q$, the functions θ_{00} and θ_{01} are interchanged, whereas the functions $q^{-1/4}\theta_{10}$ and $q^{-1/4}\theta_{11}$ are unaltered. Hence

$$\begin{aligned} \theta_{00}(v;\tau+1) &= \theta_{01}(v;\tau), \quad \theta_{10}(v;\tau+1) = e^{\pi i/4}\theta_{10}(v;\tau), \\ \theta_{01}(v;\tau+1) &= \theta_{00}(v;\tau), \quad \theta_{11}(v;\tau+1) = e^{\pi i/4}\theta_{11}(v;\tau). \end{aligned} \tag{37}$$

From Proposition 3 we obtain also the transformation formulas

$$\begin{aligned} \theta_{00}(v;-1/\tau) &= (\tau/i)^{1/2} e^{\pi i \tau v^2} \theta_{00}(\tau v;\tau), \\ \theta_{10}(v;-1/\tau) &= (\tau/i)^{1/2} e^{\pi i \tau v^2} \theta_{01}(\tau v;\tau), \\ \theta_{01}(v;-1/\tau) &= (\tau/i)^{1/2} e^{\pi i \tau v^2} \theta_{10}(\tau v;\tau), \\ \theta_{11}(v;-1/\tau) &= -i(\tau/i)^{1/2} e^{\pi i \tau v^2} \theta_{11}(\tau v;\tau). \end{aligned} \tag{38}$$

Up to this point we have used Hermite's notation just to dress up old results in new clothes. The next result breaks fresh ground.

Proposition 4 *For all $v, w \in \mathbb{C}$ and $\tau \in \mathscr{H}$,*

$$\begin{aligned} &\theta_{00}(v;\tau)\theta_{00}(w;\tau) = \theta_{00}(v+w;2\tau)\theta_{00}(v-w;2\tau) + \theta_{10}(v+w;2\tau)\theta_{10}(v-w;2\tau), \\ &\theta_{10}(v;\tau)\theta_{10}(w;\tau) = \theta_{10}(v+w;2\tau)\theta_{00}(v-w;2\tau) + \theta_{00}(v+w;2\tau)\theta_{10}(v-w;2\tau), \\ &\theta_{00}(v;\tau)\theta_{01}(w;\tau) = \theta_{01}(v+w;2\tau)\theta_{01}(v-w;2\tau) + \theta_{11}(v+w;2\tau)\theta_{11}(v-w;2\tau), \\ &\theta_{01}(v;\tau)\theta_{01}(w;\tau) = \theta_{00}(v+w;2\tau)\theta_{00}(v-w;2\tau) - \theta_{10}(v+w;2\tau)\theta_{10}(v-w;2\tau), \\ &\theta_{10}(v;\tau)\theta_{11}(w;\tau) = \theta_{11}(v+w;2\tau)\theta_{01}(v-w;2\tau) - \theta_{01}(v+w;2\tau)\theta_{11}(v-w;2\tau), \\ &\theta_{11}(v;\tau)\theta_{11}(w;\tau) = \theta_{10}(v+w;2\tau)\theta_{00}(v-w;2\tau) - \theta_{00}(v+w;2\tau)\theta_{10}(v-w;2\tau). \end{aligned}$$

Proof From the definition of θ_{00},

$$\theta_{00}(v;\tau)\theta_{00}(w;\tau) = \sum_{j,k} e^{\pi i\tau(j^2+k^2)} e^{2\pi ivj} e^{2\pi iwk} = \sum_{j+k\ even} + \sum_{j+k\ odd}.$$

In the first sum on the right we can write $j+k=2m$, $j-k=2n$. Then $j=m+n$, $k=m-n$ and

$$\begin{aligned}\sum_{j+k\ even} &= \sum_{m,n\in\mathbb{Z}} e^{2\pi i\tau(m^2+n^2)} e^{2\pi i(v+w)m} e^{2\pi i(v-w)n} \\ &= \theta_{00}(v+w;2\tau)\theta_{00}(v-w;2\tau).\end{aligned}$$

In the second sum we can write $j+k=2m+1$, $j-k=2n+1$. Then $j=m+n+1$, $k=m-n$ and

$$\begin{aligned}\sum_{j+k\ odd} &= \sum_{m,n\in\mathbb{Z}} e^{2\pi i\tau\{(m+1/2)^2+(n+1/2)^2\}} e^{2\pi iv(m+n+1)} e^{2\pi iw(m-n)} \\ &= \theta_{10}(v+w;2\tau)\theta_{10}(v-w;2\tau).\end{aligned}$$

Adding, we obtain the first relation of the proposition.

We obtain the second relation from the first by replacing v by $v+\tau/2$ and w by $w+\tau/2$. The remaining relations are obtained from the first two by increasing v and/or w by 1/2. □

By taking $w=v$ in Proposition 4, and adding or subtracting pairs of equations whose right sides differ only in one sign, we obtain the *duplication formulas*:

Proposition 5 *For all $v\in\mathbb{C}$ and $\tau\in\mathscr{H}$,*

$$\begin{aligned}\theta_{00}(2v;2\tau) &= [\theta_{00}^2(v;\tau)+\theta_{01}^2(v;\tau)]/2\theta_{00}(0;2\tau) \\ &= [\theta_{10}^2(v;\tau)-\theta_{11}^2(v;\tau)]/2\theta_{10}(0;2\tau), \\ \theta_{10}(2v;2\tau) &= [\theta_{00}^2(v;\tau)-\theta_{01}^2(v;\tau)]/2\theta_{10}(0;2\tau) \\ &= [\theta_{10}^2(v;\tau)+\theta_{11}^2(v;\tau)]/2\theta_{00}(0;2\tau), \\ \theta_{01}(2v;2\tau) &= \theta_{00}(v;\tau)\theta_{01}(v;\tau)/\theta_{01}(0;2\tau), \\ \theta_{11}(2v;2\tau) &= \theta_{10}(v;\tau)\theta_{11}(v;\tau)/\theta_{01}(0;2\tau).\end{aligned}$$

From Proposition 4 we can also derive the following *addition formulas*:

Proposition 6 *For all $v,w\in\mathbb{C}$ and $\tau\in\mathscr{H}$,*

$$\begin{aligned}&\theta_{01}^2(0)\theta_{01}(v+w)\theta_{01}(v-w) \\ &\quad = \theta_{01}^2(v)\theta_{01}^2(w)-\theta_{11}^2(v)\theta_{11}^2(w) = \theta_{00}^2(v)\theta_{00}^2(w)-\theta_{10}^2(v)\theta_{10}^2(w), \\ &\theta_{00}(0)\theta_{01}(0)\theta_{00}(v+w)\theta_{01}(v-w) \\ &\quad = \theta_{00}(v)\theta_{01}(v)\theta_{00}(w)\theta_{01}(w)+\theta_{10}(v)\theta_{11}(v)\theta_{10}(w)\theta_{11}(w), \\ &\theta_{01}(0)\theta_{10}(0)\theta_{10}(v+w)\theta_{01}(v-w) \\ &\quad = \theta_{01}(v)\theta_{10}(v)\theta_{01}(w)\theta_{10}(w)+\theta_{00}(v)\theta_{11}(v)\theta_{00}(w)\theta_{11}(w), \\ &\theta_{00}(0)\theta_{10}(0)\theta_{11}(v+w)\theta_{01}(v-w) \\ &\quad = \theta_{01}(v)\theta_{11}(v)\theta_{00}(w)\theta_{10}(w)+\theta_{00}(v)\theta_{10}(v)\theta_{01}(w)\theta_{11}(w),\end{aligned}$$

where all theta functions have the same second argument τ.

Proof Consider the second relation. If we use the first and fourth relations of Proposition 4 to evaluate the products $\theta_{00}(v)\theta_{00}(w)$ and $\theta_{01}(v)\theta_{01}(w)$, we obtain

$$\theta_{00}(v)\theta_{01}(v)\theta_{00}(w)\theta_{01}(w) = \theta_{00}^2(v+w;2\tau)\theta_{00}^2(v-w;2\tau) - \theta_{10}^2(v+w;2\tau)\theta_{10}^2(v-w;2\tau).$$

Similarly, if we use the second and sixth relations of Proposition 4 to evaluate the products $\theta_{10}(v)\theta_{10}(w)$ and $\theta_{11}(v)\theta_{11}(w)$, we obtain

$$\theta_{10}(v)\theta_{11}(v)\theta_{10}(w)\theta_{11}(w) = \theta_{10}^2(v+w;2\tau)\theta_{00}^2(v-w;2\tau) - \theta_{00}^2(v+w;2\tau)\theta_{10}^2(v-w;2\tau).$$

Hence, in the second relation of the present proposition the right side is equal to

$$[\theta_{00}^2(v+w;2\tau)+\theta_{10}^2(v+w;2\tau)][\theta_{00}^2(v-w;2\tau)-\theta_{10}^2(v-w;2\tau)].$$

On the other hand, if we use the first and fourth relations of Proposition 4 to evaluate the products $\theta_{00}(0)\theta_{00}(v+w)$ and $\theta_{01}(0)\theta_{01}(v-w)$, we see that the left side is likewise equal to

$$[\theta_{00}^2(v+w;2\tau)+\theta_{10}^2(v+w;2\tau)][\theta_{00}^2(v-w;2\tau)-\theta_{10}^2(v-w;2\tau)].$$

This proves the second relation of the proposition, and the others may be proved similarly. □

Corollary 7 *For all $v \in \mathbb{C}$ and $\tau \in \mathscr{H}$,*

$$\theta_{00}^2(0)\theta_{01}^2(v) + \theta_{10}^2(0)\theta_{11}^2(v) = \theta_{01}^2(0)\theta_{00}^2(v), \tag{39}$$

$$\theta_{10}^2(0)\theta_{01}^2(v) + \theta_{00}^2(0)\theta_{11}^2(v) = \theta_{01}^2(0)\theta_{10}^2(v). \tag{40}$$

Moreover, for all $\tau \in \mathscr{H}$,

$$\theta_{00}^4(0) = \theta_{01}^4(0) + \theta_{10}^4(0). \tag{41}$$

Proof We get (39) and (40) from the first relation of Proposition 6 by taking $w = 1/2$ and $w = (1+\tau)/2$ respectively. We obtain (41) from (39) by taking $v = 1/2$. □

If we regard (39) and (40) as a system of simultaneous linear equations for the unknowns $\theta_{01}^2(v), \theta_{11}^2(v)$, then the determinant of this system is $\theta_{00}^4(0) - \theta_{10}^4(0) = \theta_{01}^4(0) \neq 0$. It follows that the square of any theta function may be expressed as a linear combination of the squares of any other two theta functions.

By substituting for the theta functions their expansions as infinite products, the formula (41) may be given the following remarkable form:

$$\prod_{n=1}^{\infty}(1+q^{2n-1})^8 = \prod_{n=1}^{\infty}(1-q^{2n-1})^8 + 16q\prod_{n=1}^{\infty}(1+q^{2n})^8.$$

Proposition 8 *For all $v \in \mathbb{C}$ and $\tau \in \mathscr{H}$,*

$$\{\theta_{00}(v)/\theta_{01}(v)\}' = \pi i \theta_{10}^2(0)\theta_{10}(v)\theta_{11}(v)/\theta_{01}^2(v), \tag{42}$$

$$\{\theta_{10}(v)/\theta_{01}(v)\}' = \pi i \theta_{00}^2(0)\theta_{00}(v)\theta_{11}(v)/\theta_{01}^2(v), \tag{43}$$

$$\{\theta_{11}(v)/\theta_{01}(v)\}' = \pi i \theta_{01}^2(0)\theta_{00}(v)\theta_{10}(v)/\theta_{01}^2(v), \tag{44}$$

$$\{\theta_{01}'(v)/\theta_{01}(v)\}' = \theta_{01}''(0)/\theta_{01}(0) + \pi^2\theta_{00}^2(0)\theta_{10}^2(0)\theta_{11}^2(v)/\theta_{01}^2(v). \tag{45}$$

Proof By differentiating the second relation of Proposition 6 with respect to w and then putting $w = 0$, we obtain

$$\theta_{00}(0)\theta_{01}(0)[\theta_{00}'(v)\theta_{01}(v) - \theta_{00}(v)\theta_{01}'(v)] = \theta_{10}(0)\theta_{11}'(0)\theta_{10}(v)\theta_{11}(v),$$

since not only $\theta_{11}(0) = 0$ but also $\theta_{00}'(0) = \theta_{01}'(0) = \theta_{10}'(0) = 0$. Dividing by $\theta_{01}^2(v)$ and recalling the expression (36) for $\theta_{11}'(0)$, we obtain (42). Similarly, from the third and fourth relations of Proposition 6 we obtain (43) and (44).

In the same way, if we differentiate the first relation of Proposition 6 twice with respect to w and then put $w = 0$, we obtain

$$\theta_{01}^2(0)[\theta_{01}''(v)\theta_{01}(v) - \theta_{01}'(v)^2] = \theta_{01}(0)\theta_{01}''(0)\theta_{01}^2(v) - \theta_{11}'(0)^2\theta_{11}^2(v).$$

Hence, using (36) again, we obtain (45). □

We are now in a position to make the connection between theta functions and elliptic functions.

5 Jacobian Elliptic Functions

The behaviour of the theta functions when their argument is increased by 1 or τ makes it clear that doubly-periodic functions may be constructed from their quotients. We put

$$\begin{aligned} \operatorname{sn} u = \operatorname{sn}(u;\tau) &:= -i\theta_{00}(0)\theta_{11}(v)/\theta_{10}(0)\theta_{01}(v), \\ \operatorname{cn} u = \operatorname{cn}(u;\tau) &:= \theta_{01}(0)\theta_{10}(v)/\theta_{10}(0)\theta_{01}(v), \\ \operatorname{dn} u = \operatorname{dn}(u;\tau) &:= \theta_{01}(0)\theta_{00}(v)/\theta_{00}(0)\theta_{01}(v), \end{aligned} \tag{46}$$

where $u = \pi\theta_{00}^2(0)v$.

The constant multiples are chosen so that, in addition to $\operatorname{sn} 0 = 0$, we have $\operatorname{cn} 0 = \operatorname{dn} 0 = 1$. The independent variable is scaled so that, by (42)–(44),

$$\begin{aligned} d(\operatorname{sn} u)/du &= \operatorname{cn} u \operatorname{dn} u, \\ d(\operatorname{cn} u)/du &= -\operatorname{sn} u \operatorname{dn} u, \\ d(\operatorname{dn} u)/du &= -\lambda \operatorname{sn} u \operatorname{cn} u, \end{aligned} \tag{47}$$

where

$$\lambda = \lambda(\tau) := \theta_{10}^4(0;\tau)/\theta_{00}^4(0;\tau). \tag{48}$$

It follows at once from the definitions that $\operatorname{sn} u$ is an odd function of u, whereas $\operatorname{cn} u$ and $\operatorname{dn} u$ are even functions of u. It follows from (41) that

$$1 - \lambda(\tau) = \theta_{01}^4(0;\tau)/\theta_{00}^4(0;\tau), \tag{49}$$

and from (39)–(40) that

$$\mathrm{cn}^2 u = 1 - \mathrm{sn}^2 u, \quad \mathrm{dn}^2 u = 1 - \lambda \mathrm{sn}^2 u. \tag{50}$$

Evidently (47) implies

$$d(\mathrm{sn}^2 u)/du = 2\mathrm{sn}\, u\, \mathrm{cn}\, u\, \mathrm{dn}\, u,$$
$$d^2(\mathrm{sn}^2 u)/du^2 = 2(\mathrm{cn}^2 u\, \mathrm{dn}^2 u - \mathrm{sn}^2 u\, \mathrm{dn}^2 u - \lambda \mathrm{sn}^2 u\, \mathrm{cn}^2 u).$$

If we write $S(u) = S(u;\tau) := \mathrm{sn}^2 u$ and use (50), we can rewrite this in the form

$$d^2 S/du^2 = 2[(1-S)(1-\lambda S) - S(1-\lambda S) - \lambda S(1-S)]$$
$$= 6\lambda S^2 - 4(1+\lambda)S + 2.$$

Since $S(0) = S'(0) = 0$, we conclude that $S(u)$ coincides with the function denoted by the same symbol in §3. However, it should be noted that now λ is not given, but is determined by τ. Thus the question arises: can we choose $\tau \in \mathscr{H}$ (the upper half-plane) so that $\lambda(\tau)$ is any prescribed complex number other than 0 or 1?

For many applications it is sufficient to know that we can choose $\tau \in \mathscr{H}$ so that $\lambda(\tau)$ is any prescribed real number between 0 and 1. Since this case is much simpler, we will deal with it now and defer treatment of the general case until the next section. We have

$$\lambda(\tau) = 1 - \theta_{01}^4(0;\tau)/\theta_{00}^4(0;\tau) = 1 - \prod_{n=1}^{\infty}\{(1-q^{2n-1})/(1+q^{2n-1})\}^8,$$

where $q = e^{\pi i \tau}$. If $\tau = iy$, where $y > 0$, then $0 < q < 1$. Moreover, as y increases from 0 to ∞, q decreases from 1 to 0 and the infinite product increases from 0 to 1. Thus $\lambda(\tau)$ decreases continuously from 1 to 0 and, for each $w \in (0,1)$, there is a unique pure imaginary $\tau \in \mathscr{H}$ such that $\lambda(\tau) = w$.

It should be mentioned that, also with our previous approach, $S(u)$ could have been recognized as the square of a meromorphic function by defining $\mathrm{sn}\, u$, $\mathrm{cn}\, u$, $\mathrm{dn}\, u$ to be the solution, for *given* $\lambda \in \mathbb{C}$, of the system of differential equations (47) which satisfies the initial condition $\mathrm{sn}\, 0 = 0$, $\mathrm{cn}\, 0 = \mathrm{dn}\, 0 = 1$.

Elliptic functions were first defined by Abel (1827) as the inverses of elliptic integrals. His definitions were modified by Jacobi (1829) to accord with Legendre's normal form for elliptic integrals, and the functions $\mathrm{sn}\, u$, $\mathrm{cn}\, u$, $\mathrm{dn}\, u$ are generally known as the *Jacobian elliptic functions*. The actual notation is due to Gudermann (1838). The definition by means of theta functions was given later by Jacobi (1838) in lectures.

Several properties of the Jacobian elliptic functions are easy consequences of the later definition. In the first place, all three are meromorphic in the whole u-plane, since the theta functions are everywhere holomorphic. Their poles are determined by the zeros of $\theta_{01}(v)$ and are all simple. Similarly, the zeros of $\mathrm{sn}\, u$, $\mathrm{cn}\, u$ and $\mathrm{dn}\, u$ are determined by the zeros of $\theta_{11}(v)$, $\theta_{10}(v)$ and $\theta_{00}(v)$ respectively and are all simple. If we put

$$\mathbf{K} = \mathbf{K}(\tau) := \pi\theta_{00}^2(0;\tau)/2, \quad \mathbf{K}' = \mathbf{K}'(\tau) := \tau\mathbf{K}(\tau)/i, \tag{51}$$

then we have

$$\begin{aligned} \textit{Poles of } \operatorname{sn} u, \operatorname{cn} u, \operatorname{dn} u: &\quad u = 2m\mathbf{K} + (2n+1)i\mathbf{K}' \ (m,n \in \mathbb{Z}). \qquad (52)\\ \textit{Zeros of } \operatorname{sn} u: &\quad u = 2m\mathbf{K} + 2ni\mathbf{K}', \\ \operatorname{cn} u: &\quad u = (2m+1)\mathbf{K} + 2ni\mathbf{K}', \ (m,n \in \mathbb{Z}) \qquad (53)\\ \operatorname{dn} u: &\quad u = (2m+1)\mathbf{K} + (2n+1)i\mathbf{K}'. \end{aligned}$$

From the definitions (46) of the Jacobian elliptic functions and the behaviour of the theta functions when v is increased by 1 or τ we further obtain

$$\begin{aligned} \operatorname{sn} u &= -\operatorname{sn}(u + 2\mathbf{K}) = \operatorname{sn}(u + 2i\mathbf{K}'),\\ \operatorname{cn} u &= -\operatorname{cn}(u + 2\mathbf{K}) = -\operatorname{cn}(u + 2i\mathbf{K}'),\\ \operatorname{dn} u &= \operatorname{dn}(u + 2\mathbf{K}) = -\operatorname{dn}(u + 2i\mathbf{K}'). \end{aligned} \tag{54}$$

It follows that all three functions are *doubly-periodic*. In fact sn u has periods $4\mathbf{K}$ and $2i\mathbf{K}'$, cn u has periods $4\mathbf{K}$ and $2\mathbf{K} + 2i\mathbf{K}'$, and dn u has periods $2\mathbf{K}$ and $4i\mathbf{K}'$. In each case the ratio of the two periods is not real, since $\tau \in \mathscr{H}$.

Since any period must equal a difference between two poles, it must have the form $2m\mathbf{K} + 2ni\mathbf{K}'$ for some $m, n \in \mathbb{Z}$. Since $4\mathbf{K}$ and $2i\mathbf{K}'$ are periods of sn u, but $2\mathbf{K}$ is not, and since any integral linear combination of periods is again a period, it follows that the periods of sn u are precisely the integral linear combinations of $4\mathbf{K}$ and $2i\mathbf{K}'$. Similarly the periods of cn u are the integral linear combinations of $4\mathbf{K}$ and $2\mathbf{K} + 2i\mathbf{K}'$, and the periods of dn u are the integral linear combinations of $2\mathbf{K}$ and $4i\mathbf{K}'$.

It was shown in §3 that, if $0 < \lambda < 1$, then $S(t, \lambda)$ has least positive period $2K(\lambda)$, where

$$K(\lambda) = \int_0^1 dx/g_\lambda(x)^{1/2}.$$

But, as we have seen, there is a unique pure imaginary $\tau \in \mathscr{H}$ such that $\lambda = \lambda(\tau)$, and $2K[\lambda(\tau)]$ is then the least positive period of $\operatorname{sn}^2(u;\tau)$. Since the periods of $\operatorname{sn}^2(u;\tau)$ are $2m\mathbf{K} + 2ni\mathbf{K}'(m, n \in \mathbb{Z})$, and since $\mathbf{K}, \mathbf{K}'$ are real and positive when τ is pure imaginary, it follows that

$$K[\lambda(\tau)] = \mathbf{K}(\tau).$$

The domain of validity of this relation may be extended by appealing to results which will be established in §6. In fact it holds, by analytic continuation, for all τ in the region $\mathscr{D}$ illustrated in Figure 3, since $\lambda(\tau) \in \mathscr{H}$ for $\tau \in \mathscr{D}$.

From the definitions (46) of the Jacobian elliptic functions, the addition formulas for the theta functions (Proposition 6) and the expression (48) for λ, we obtain *addition formulas* for the Jacobian functions:

$$\begin{aligned} \operatorname{sn}(u_1+u_2) &= (\operatorname{sn}u_1\operatorname{cn}u_2\operatorname{dn}u_2 + \operatorname{sn}u_2\operatorname{cn}u_1\operatorname{dn}u_1)/(1-\lambda\operatorname{sn}^2u_1\operatorname{sn}^2u_2),\\ \operatorname{cn}(u_1+u_2) &= (\operatorname{cn}u_1\operatorname{cn}u_2 - \operatorname{sn}u_1\operatorname{sn}u_2\operatorname{dn}u_1\operatorname{dn}u_2)/(1-\lambda\operatorname{sn}^2u_1\operatorname{sn}^2u_2), \qquad (55)\\ \operatorname{dn}(u_1+u_2) &= (\operatorname{dn}u_1\operatorname{dn}u_2 - \lambda\operatorname{sn}u_1\operatorname{sn}u_2\operatorname{cn}u_1\operatorname{cn}u_2)/(1-\lambda\operatorname{sn}^2u_1\operatorname{sn}^2u_2). \end{aligned}$$

The addition formulas show that the evaluation of the Jacobian elliptic functions for arbitrary complex argument may be reduced to their evaluation for real and pure imaginary arguments.

The usual addition formulas for the sine and cosine functions may be regarded as limiting cases of (55). For if $\tau = iy$ and $y \to \infty$, the product expansions (35) show that

$$\theta_{00}(v) \to 1, \quad \theta_{01}(v) \to 1,$$
$$\theta_{10}(v) \sim 2e^{\pi i\tau/4}\cos \pi v, \quad \theta_{11}(v) \sim 2ie^{\pi i\tau/4}\sin \pi v,$$

and hence

$$\lambda \to 0, \quad u \to \pi v,$$
$$\operatorname{sn} u \to \sin u, \quad \operatorname{cn} u \to \cos u, \quad \operatorname{dn} u \to 1.$$

The definitions (46) of the Jacobian elliptic functions and the transformation formulas (37)–(38) for the theta functions imply also *transformation formulas* for the Jacobian functions:

Proposition 9 *For all $u \in \mathbb{C}$ and $\tau \in \mathscr{H}$,*

$$\operatorname{sn}(u;\tau+1) = (1-\lambda(\tau))^{1/2}\operatorname{sn}(u';\tau)/\operatorname{dn}(u';\tau),$$
$$\operatorname{cn}(u;\tau+1) = \operatorname{cn}(u';\tau)/\operatorname{dn}(u';\tau),$$
$$\operatorname{dn}(u;\tau+1) = 1/\operatorname{dn}(u';\tau),$$

where

$$u' = u/(1-\lambda(\tau))^{1/2}$$

and

$$(1-\lambda(\tau))^{1/2} = \theta_{01}^2(0;\tau)/\theta_{00}^2(0;\tau).$$

Furthermore,

$$\lambda(\tau+1) = \lambda(\tau)/[\lambda(\tau)-1],$$
$$\mathbf{K}(\tau+1) = (1-\lambda(\tau))^{1/2}\mathbf{K}(\tau).$$

Proof With $v = u/\pi\theta_{00}^2(0;\tau+1)$ we have, by (37),

$$\operatorname{dn}(u;\tau+1) = \theta_{00}(0;\tau)\theta_{01}(v;\tau)/\theta_{01}(0;\tau)\theta_{00}(v;\tau) = 1/\operatorname{dn}(u';\tau),$$

where

$$u' = \pi\theta_{00}^2(0;\tau)v = \theta_{00}^2(0;\tau)u/\theta_{01}^2(0;\tau) = u/(1-\lambda(\tau))^{1/2}.$$

Similarly, from (37) and (48)-(49), we obtain

$$\lambda(\tau+1) = -\theta_{10}^4(0;\tau)/\theta_{01}^4(0;\tau) = \lambda(\tau)/[\lambda(\tau)-1].$$

The other relations are established in the same way. □

Proposition 10 *For all $u \in \mathbb{C}$ and $\tau \in \mathscr{H}$,*

$$\begin{aligned}
\mathrm{sn}\,(u; -1/\tau) &= -i\,\mathrm{sn}\,(iu; \tau)/\mathrm{cn}\,(iu; \tau),\\
\mathrm{cn}\,(u; -1/\tau) &= 1/\mathrm{cn}(iu; \tau),\\
\mathrm{dn}\,(u; -1/\tau) &= \mathrm{dn}(iu; \tau)/cn(iu; \tau),
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\lambda(-1/\tau) &= 1 - \lambda(\tau),\\
\mathbf{K}(-1/\tau) &= \mathbf{K}'(\tau).
\end{aligned}$$

Proof With $v = u/\pi\theta_{00}^2(0; -1/\tau)$ we have, by (38),

$$\begin{aligned}
\mathrm{sn}\,(u; -1/\tau) &= -i\theta_{00}(0; -1/\tau)\theta_{11}(v; -1/\tau)/\theta_{10}(0; -1/\tau)\theta_{01}(v; -1/\tau)\\
&= -\theta_{00}(0; \tau)\theta_{11}(\tau v; \tau)/\theta_{01}(0; \tau)\theta_{10}(\tau v; \tau).
\end{aligned}$$

On the other hand, with $v' = iu/\pi\theta_{00}^2(0; \tau)$ we have

$$\mathrm{sn}\,(iu; \tau)/\mathrm{cn}\,(iu; \tau) = -i\theta_{00}(0; \tau)\theta_{11}(v'; \tau)/\theta_{01}(0; \tau)\theta_{10}(v'; \tau).$$

Since $\tau v = v'$, by comparing these two relations we obtain the first assertion of the proposition.

The next two assertions may be obtained in the same way. The final two assertions follow from (38), together with (48), (49) and (51). □

It follows from Proposition 10 that the evaluation of the Jacobian elliptic functions for pure imaginary argument and parameter τ may be reduced to their evaluation for real argument and parameter $-1/\tau$.

From the definition (46) of the Jacobian elliptic functions and the duplication formulas for the theta functions we can also obtain formulas for the Jacobian functions when the parameter τ is doubled ('Landen's transformation'):

Proposition 11 *For all $u \in \mathbb{C}$ and $\tau \in \mathscr{H}$,*

$$\begin{aligned}
\mathrm{sn}\,(u''; 2\tau) &= [1 + (1 - \lambda(\tau))^{1/2}]\mathrm{sn}\,(u; \tau)\mathrm{cn}\,(u; \tau)/\mathrm{dn}\,(u; \tau),\\
\mathrm{cn}\,(u''; 2\tau) &= \{1 - [1 + (1 - \lambda(\tau))^{1/2}]\mathrm{sn}^2(u; \tau)\}/\mathrm{dn}\,(u; \tau),\\
\mathrm{dn}\,(u''; 2\tau) &= \{1 - [1 - (1 - \lambda(\tau))^{1/2}]\mathrm{sn}^2(u; \tau)\}/\mathrm{dn}\,(u; \tau),
\end{aligned}$$

where $u'' = [1 + (1 - \lambda(\tau))^{1/2}]u$ and $(1 - \lambda(\tau))^{1/2} = \theta_{01}^2(0; \tau)/\theta_{00}^2(0; \tau)$.

Furthermore,

$$\begin{aligned}
\lambda(2\tau) &= \lambda^2(\tau)/[1 + (1 - \lambda(\tau))^{1/2}]^4,\\
\mathbf{K}(2\tau) &= [1 + (1 - \lambda(\tau))^{1/2}]\mathbf{K}(\tau)/2.
\end{aligned}$$

Proof If $u = \pi\theta_{00}^2(0; \tau)v$ and $u'' = \pi\theta_{00}^2(0; 2\tau)2v$ then, by Proposition 5,

$$\begin{aligned}
u'' &= 2\theta_{00}^2(0; 2\tau)u/\theta_{00}^2(0; \tau)\\
&= [\theta_{00}^2(0; \tau) + \theta_{01}^2(0; \tau)]u/\theta_{00}^2(0; \tau).
\end{aligned}$$

Hence, by (49),

$$u'' = [1 + (1 - \lambda(\tau))^{1/2}]u.$$

By Proposition 5 also,

$$\operatorname{sn}(u''; 2\tau) = -i\theta_{00}(0; 2\tau)\theta_{10}(v; \tau)\theta_{11}(v; \tau)/\theta_{10}(0; 2\tau)\theta_{00}(v; \tau)\theta_{01}(v; \tau).$$

On the other hand,

$$\operatorname{sn}(u; \tau)\operatorname{cn}(u; \tau)/\operatorname{dn}(u; \tau) = -i\theta_{00}^2(0; \tau)\theta_{10}(v; \tau)\theta_{11}(v; \tau)/D,$$

where $D = \theta_{10}^2(0; \tau)\theta_{00}(v; \tau)\theta_{01}(v; \tau)$.

Since $2\theta_{00}(0; 2\tau)\theta_{10}(0; 2\tau) = \theta_{10}^2(0; \tau)$, it follows that

$$\operatorname{sn}(u''; 2\tau) = 2\theta_{00}^2(0; 2\tau)\operatorname{sn}(u; \tau)\operatorname{cn}(u; \tau)/\theta_{00}^2(0; \tau)\operatorname{dn}(u; \tau).$$

Since $2\theta_{00}^2(0; 2\tau)/\theta_{00}^2(0; \tau) = u''/u$, this proves the first assertion of the proposition. The remaining assertions may be proved similarly. □

We show finally how the standard elliptic integrals of the second and third kinds, defined by (24) and (25), may be expressed in terms of theta functions. If we put

$$\Theta(u) = \theta_{01}(v), \tag{56}$$

where $u = \pi\theta_{00}^2(0)v$, then since

$$\lambda S(u) = \lambda \operatorname{sn}^2 u = -\theta_{10}^2(0)\theta_{11}^2(v)/\theta_{00}^2(0)\theta_{01}^2(v),$$

we can rewrite (45) in the form

$$d\{\Theta'(u)/\Theta(u)\}/du = -\alpha + 1 - \lambda S(u),$$

where α is independent of u and the prime on the left denotes differentiation with respect to u. Since $\Theta'(0) = 0$, by integrating we obtain

$$E(u) = \Theta'(u)/\Theta(u) + \alpha u.$$

To determine α we take $u = \mathbf{K}$. Since $\theta_{01}'(1/2) = \theta_{00}'(1) = \theta_{00}'(0) = 0$, we obtain $\alpha = \mathbf{E}/\mathbf{K}$, where

$$\mathbf{E} = \mathbf{E}(\mathbf{K}) = \int_0^K \{1 - \lambda S(u)\}\, du = \int_0^1 (1 - \lambda x)\, dx/g_\lambda(x)^{1/2}$$

is a complete elliptic integral of the second kind. Thus

$$E(u) = \Theta'(u)/\Theta(u) + u\mathbf{E}/\mathbf{K}. \tag{57}$$

Substituting this expression for $E(u)$ in (27), we further obtain

$$\Pi(u, a) = u\Theta'(a)/\Theta(a) + (1/2)\log\{\Theta(u - a)/\Theta(u + a)\}. \tag{58}$$

6 The Modular Function

The function

$$\lambda(\tau) := \theta_{10}^4(0;\tau)/\theta_{00}^4(0;\tau),$$

which was introduced in §5, is known as the *modular function*. In this section we study its remarkable properties. (The term 'modular function', without the definite article, is also used in a more general sense, which we do not consider here.)

The modular function is holomorphic in the upper half-plane $\mathscr{H}$. Furthermore, we have

Proposition 12 *For any $\tau \in \mathscr{H}$,*

$$\begin{aligned}
\lambda(\tau+1) &= \lambda(\tau)/[\lambda(\tau)-1],\\
\lambda(-1/\tau) &= 1-\lambda(\tau),\\
\lambda(-1/(\tau+1)) &= 1/[1-\lambda(\tau)],\\
\lambda((\tau-1)/\tau) &= [\lambda(\tau)-1]/\lambda(\tau),\\
\lambda(\tau/(\tau+1)) &= 1/\lambda(\tau).
\end{aligned}$$

Proof The first two relations have already been established in Propositions 9 and 10. If, as in §1, we put

$$U\lambda = 1-\lambda, \quad V\lambda = 1/(1-\lambda),$$

and if we also put $T\tau = \tau+1$, $S\tau = -1/\tau$, then they may be written in the form

$$\lambda(T\tau) = UV\lambda(\tau), \quad \lambda(S\tau) = U\lambda(\tau).$$

It follows that

$$\lambda(-1/(\tau+1)) = \lambda(ST\tau) = U\lambda(T\tau) = U^2V\lambda(\tau) = V\lambda(\tau) = 1/[1-\lambda(\tau)].$$

Similarly,

$$\begin{aligned}
\lambda((\tau-1)/\tau) &= \lambda(TS\tau) = V^2\lambda(\tau) = [\lambda(\tau)-1]/\lambda(\tau),\\
\lambda(\tau/(\tau+1)) &= \lambda(TST\tau) = UV^2\lambda(\tau) = 1/\lambda(\tau).
\end{aligned}$$

□

As we saw in Proposition IV.12, together the transformations $S\tau = -1/\tau$ and $T\tau = \tau+1$ generate the *modular group* Γ, consisting of all linear fractional transformations

$$\tau' = (a\tau+b)/(c\tau+d),$$

where $a, b, c, d \in \mathbb{Z}$ and $ad-bc = 1$. Consequently we can deduce the effect on $\lambda(\tau)$ of any modular transformation on τ. However, Proposition 12 contains the only cases which we require.

We will now study in some detail the behaviour of the modular function in the upper half-plane. We first observe that we need only consider the behaviour of $\lambda(\tau)$ in the right half of $\mathscr{H}$. For, from the definitions of the theta functions as infinite series,

$$\overline{\theta_{00}(0;\tau)} = \theta_{00}(0;-\bar{\tau}), \quad \overline{\theta_{01}(0;\tau)} = \theta_{01}(0;-\bar{\tau}),$$

where the bar denotes complex conjugation, and hence

$$\lambda(-\bar{\tau}) = \overline{\lambda(\tau)}. \tag{59}$$

We next note that, by taking $\tau = i$ in the relation $\lambda(-1/\tau) = 1 - \lambda(\tau)$, we obtain $\lambda(i) = 1/2$. We have already seen in §5 that $\lambda(\tau)$ is real on the imaginary axis $\tau = iy$ $(y > 0)$, and decreases from 1 to 0 as y increases from 0 to ∞. Since $\lambda(\tau + 1) = \lambda(\tau)/[\lambda(\tau) - 1]$, it follows that $\lambda(\tau)$ is real also on the half-line $\tau = 1 + iy$ $(y > 0)$, and increases from $-\infty$ to 0 as y increases from 0 to ∞. Moreover, $\lambda(1 + i) = -1$.

The linear fractional map $\tau = (\tau' - 1)/\tau'$ maps the half-line $\mathscr{R}\tau' = 1$, $\mathscr{I}\tau' > 0$ onto the semi-circle $|\tau - 1/2| = 1/2$, $\mathscr{I}\tau > 0$, and $\tau' = 1 + i$ is mapped to $\tau = (1 + i)/2$. Since

$$\lambda((\tau' - 1)/\tau') = [\lambda(\tau') - 1]/\lambda(\tau'),$$

it follows from what we have just proved that, as τ traverses this semi-circle from 0 to 1, $\lambda(\tau)$ is real and increases from 1 to ∞. Moreover, $\lambda((1 + i)/2) = 2$.

If $\mathscr{R}\tau = 1/2$, then $\bar{\tau} = 1 - \tau$ and hence, by (59),

$$\overline{\lambda(\tau)} = \lambda(\tau - 1) = \lambda(\tau)/[\lambda(\tau) - 1],$$

which implies

$$|\lambda(\tau) - 1|^2 = 1.$$

Thus $w = \lambda(\tau)$ maps the half-line $\mathscr{R}\tau = 1/2$, $\mathscr{I}\tau > 0$ into the circle $|w - 1| = 1$. Furthermore, the map is injective. For if $\lambda(\tau_1) = \lambda(\tau_2)$, then $\lambda(2\tau_1) = \lambda(2\tau_2)$, by Proposition 11, and the map is injective on the half-line $\mathscr{R}\tau = 1$, $\mathscr{I}\tau > 0$. If $\tau = 1/2 + iy$, where $y \to +\infty$, then

$$\theta_{00}(0; \tau) \to 1, \quad \theta_{10}(0; \tau) \sim 2e^{\pi i\tau/4}$$

and hence

$$\lambda(\tau) \sim 16ie^{-\pi y}.$$

In particular, $\lambda(\tau) \in \mathscr{H}$ and $\lambda(\tau) \to 0$. Since $\lambda((1 + i)/2) = 2$, it follows that $w = \lambda(\tau)$ maps the half-line $\tau = 1/2 + iy$ $(y > 1/2)$ bijectively onto the semi-circle $|w - 1| = 1$, $\mathscr{I}w > 0$.

If $|\tau| = 1$, $\mathscr{I}\tau > 0$ and $\tau' = \tau/(1 + \tau)$, then $\mathscr{R}\tau' = 1/2$, $\mathscr{I}\tau' > 0$ and $\lambda(\tau') = 1/\lambda(\tau)$. Consequently, by what we have just proved, $w = \lambda(\tau)$ maps the semi-circle $|\tau| = 1$, $\mathscr{I}\tau > 0$ bijectively onto the half-line $\mathscr{R}w = 1/2$, $\mathscr{I}w > 0$.

The point $e^{\pi i/3} = (1 + i\sqrt{3})/2$ is in $\mathscr{H}$ and lies on both the line $\mathscr{R}\tau = 1/2$ and the circle $|\tau| = 1$. Hence $\lambda(e^{\pi i/3})$ lies on both the semi-circle $|w - 1| = 1$, $\mathscr{I}w > 0$ and the line $\mathscr{R}w = 1/2$, which implies that

$$\lambda(e^{\pi i/3}) = e^{\pi i/3}.$$

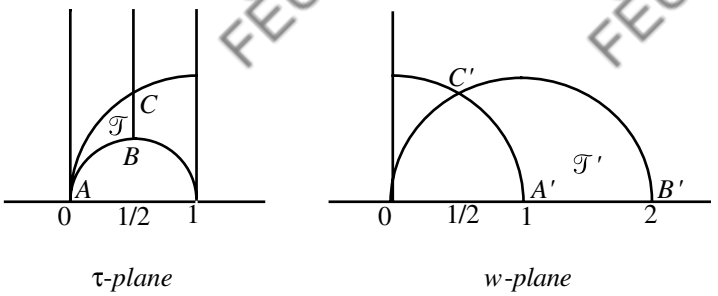

Fig. 2. $w = \lambda(\tau)$ maps $\mathscr{T}$ onto $\mathscr{T}'$.

Again, since $\lambda(\tau - 1) = \lambda(\tau)/[\lambda(\tau) - 1]$, $w = \lambda(\tau)$ maps the semi-circle $|\tau - 1| = 1$, $\mathscr{I}\tau > 0$ bijectively onto the semi-circle $|w| = 1$, $\mathscr{I}w > 0$.

In particular, we have the behaviour illustrated in Figure 2: $w = \lambda(\tau)$ maps the boundary of the (non-Euclidean) 'triangle' $\mathscr{T}$ with vertices $A = 0$, $B = (1 + i)/2$, $C = e^{\pi i/3}$ bijectively onto the boundary of the 'triangle' $\mathscr{T}'$ with vertices $A' = 1$, $B' = 2$, $C' = e^{\pi i/3}$. We are going to deduce from this that the region inside $\mathscr{T}$ is mapped bijectively onto the region inside $\mathscr{T}'$. The reasoning here does not depend on special properties of the function or the domain, but is quite general (the 'principle of the argument'). To emphasize this, we will temporarily denote the independent variable by z, instead of τ.

Choose any $w_0 \in \mathbb{C}$ which is either inside or outside the 'triangle' $\mathscr{T}'$, and let Δ denote the change in the argument of $w - w_0$ as w traverses $\mathscr{T}'$ in the direction $A'B'C'$. Thus $\Delta = 2\pi$ or 0 according as w_0 is inside or outside $\mathscr{T}'$. But Δ is also the change in the argument of $\lambda(z) - w_0$ as z traverses $\mathscr{T}$ in the direction ABC. Since $\lambda(z)$ is a nonconstant holomorphic function, the number of times that it assumes the value w_0 inside $\mathscr{T}$ is either zero or a positive integer p.

Suppose the latter, and let $z = \zeta_1, \ldots, \zeta_p$ be the points inside $\mathscr{T}$ for which $\lambda(z) = w_0$. In the neighbourhood of ζ_j we have, for some positive integer m_j and some $a_{0j} \neq 0$,

$$\lambda(z) - w_0 = a_{0j}(z - \zeta_j)^{m_j} + a_{1j}(z - \zeta_j)^{m_j+1} + \cdots$$

and

$$\lambda'(z) = m_j a_{0j}(z - \zeta_j)^{m_j-1} + (m_j + 1)a_{1j}(z - \zeta_j)^{m_j} + \cdots .$$

Hence

$$\lambda'(z)/[\lambda(z) - w_0] = m_j/(z - \zeta_j) + f_j(z),$$

where $f_j(z)$ is holomorphic at ζ_j. Consequently

$$f(z) := \lambda'(z)/[\lambda(z) - w_0] - \sum_{j=1}^{p} m_j/(z - \zeta_j)$$

is holomorphic at every point z inside $\mathscr{T}$. Hence, by Cauchy's theorem,

$$\int_{\mathscr{T}} f(z)\,dz = 0.$$

But, since $\log \lambda(z) = \log|\lambda(z)| + i \arg \lambda(z)$,

$$\int_{\mathscr{T}} \lambda'(z)\,dz/[\lambda(z) - w_0] = i\,\Delta.$$

Similarly, since ζ_j is inside $\mathscr{T}$,

$$\int_{\mathscr{T}} dz/(z - \zeta_j) = 2\pi i.$$

It follows that

$$\Delta = 2\pi \sum_{j=1}^{p} m_j.$$

If w_0 is outside $\mathscr{T}'$, then $\Delta = 0$ and we have a contradiction. Hence $\lambda(z)$ is never outside $\mathscr{T}'$ if z is inside $\mathscr{T}$. If w_0 is inside $\mathscr{T}'$, then $\Delta = 2\pi$. Hence $\lambda(z)$ assumes each value inside $\mathscr{T}'$ at exactly one point z inside $\mathscr{T}$, and at this point $\lambda'(z) \neq 0$.

Finally, if $\lambda(z)$ assumed a value w_0 *on* $\mathscr{T}'$ at a point z_0 inside $\mathscr{T}$, then it would assume all values near w_0 in the neighbourhood of z_0. In particular, it would assume values outside $\mathscr{T}'$, which we have shown to be impossible. It follows that $w = \lambda(z)$ maps the region inside $\mathscr{T}$ bijectively onto the region inside $\mathscr{T}'$, and $\lambda'(z) \neq 0$ for all z inside $\mathscr{T}$.

We must also have $\lambda'(z) \neq 0$ for all $z \neq 0$ on $\mathscr{T}$. Otherwise, if $\lambda(z_0) = w_0$ and $\lambda'(z_0) = 0$ for some $z_0 \in \mathscr{T} \cap \mathscr{H}$ then, for some $m > 1$ and $c \neq 0$,

$$\lambda(z) - w_0 \sim c(z - z_0)^m \text{ as } z \to z_0.$$

But this implies that $\lambda(z)$ takes values outside $\mathscr{T}'$ for some z near z_0 inside $\mathscr{T}$.

By putting together the preceding results we see that $w = \lambda(\tau)$ maps the domain

$$\mathscr{D} = \{\tau \in \mathscr{H} : 0 < \mathscr{R}\tau < 1, |\tau - 1/2| > 1/2\}$$

bijectively onto the upper half-plane $\mathscr{H}$, with the subdomain k of $\mathscr{D}$ mapped onto the subdomain k' of $\mathscr{H}$ $(k = 1, \ldots, 6)$, as illustrated in Figure 3. Moreover, the boundary in $\mathscr{H}$ of $\mathscr{D}$ is mapped bijectively onto the real axis, with the points 0 and 1 omitted.

If we denote by $\bar{\mathscr{D}}$ the closure of $\mathscr{D}$ in $\mathscr{H}$ and by $\mathscr{D}^*$ the reflection of $\mathscr{D}$ in the imaginary axis, then it follows from (59) that $w = \lambda(\tau)$ maps the region

$$\begin{aligned}\bar{\mathscr{D}} \cup \mathscr{D}^* = {} & \{\tau \in \mathscr{H} : 0 \leq \mathscr{R}\tau \leq 1, |\tau - 1/2| \geq 1/2\} \\ & \cup \{\tau \in \mathscr{H} : -1 < \mathscr{R}\tau < 0, |\tau + 1/2| > 1/2\}\end{aligned}$$

bijectively onto the whole complex plane $\mathbb{C}$, with the points 0 and 1 omitted. *This answers the question raised in* §5.

There remains the practical problem, for a given $w \in \mathbb{C}$, of determining $\tau \in \mathscr{H}$ such that $\lambda(\tau) = w$. If $0 < w < 1$, we can calculate τ by the *AGM* algorithm, using the formula (4), since $\tau = iK(1-w)/K(w)$. For complex w we can use an extension of the *AGM* algorithm, or proceed in the following way.

Since

$$(1-\lambda(\tau))^{1/4} = \theta_{01}(0;\tau)/\theta_{00}(0;\tau)$$

and

$$\theta_{00}(0;\tau) = 1 + 2\sum_{n=1}^{\infty} q^{n^2}, \quad \theta_{01}(0;\tau) = 1 + 2\sum_{n=1}^{\infty} (-1)^n q^{n^2},$$

we have

$$\begin{aligned}&[1-(1-\lambda(\tau))^{1/4}]/[1+(1-\lambda(\tau))^{1/4}]\\ &\quad = [\theta_{00}(0;\tau) - \theta_{01}(0;\tau)]/[\theta_{00}(0;\tau) + \theta_{01}(0;\tau)]\\ &\quad = 2(q + q^9 + q^{25} + \cdots)/(1 + 2q^4 + 2q^{16} + \cdots).\end{aligned}$$

Thus if we put

$$\ell := [1-(1-w)^{1/4}]/[1+(1-w)^{1/4}],$$

we have to solve for q the equation

$$\ell/2 = (q + q^9 + q^{25} + \cdots)/(1 + 2q^4 + 2q^{16} + \cdots).$$

Expanding the right side as a power series in q and inverting the relationship, we obtain

$$q = \ell/2 + 2(\ell/2)^5 + 15(\ell/2)^9 + 150(\ell/2)^{13} + O(\ell/2)^{17}.$$

To ensure rapid convergence we may suppose that, in Figure 3, w is situated in the region $5'$ or on its boundary, since the general case may be reduced to this by a linear fractional transformation. It is not difficult to show that in this region $|\ell|$ takes its maximum value when $w = e^{\pi i/3}$, and then

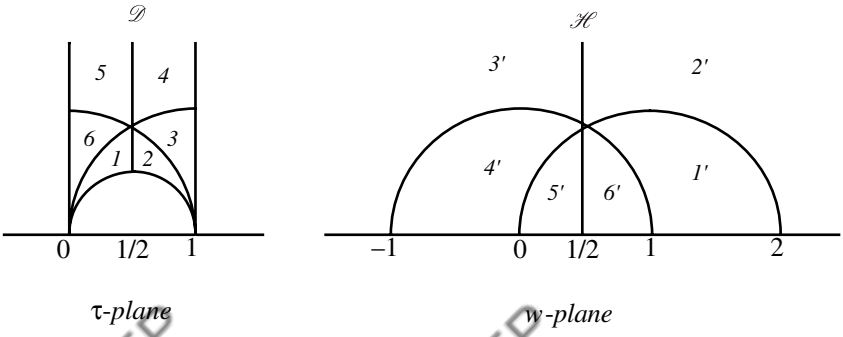

Fig. 3. $w = \lambda(\tau)$ maps $\mathscr{D}$ onto $\mathscr{H}$.

$$\ell = (1 - e^{-\pi i/12})/(1 + e^{-\pi i/12}) = i \tan \pi/24.$$

Thus $|\ell| \leq \tan \pi/24 < 2/15$ and $|\ell/2|^4 < 2 \times 10^{-5}$. Since $\mathscr{I}\tau \geq \sqrt{3}/2$ for τ in the region 5, for the solution q we have

$$|q| \leq e^{-\pi\sqrt{3}/2} < 1/15.$$

Having determined q, we may calculate $\mathbf{K}(\tau)$, sn u, ... from their representations by theta functions.

7 Further Remarks

Numerous references to the older literature on elliptic integrals and elliptic functions are given by Fricke [12]. The more important original contributions are readily available in Euler [10], Lagrange [21], Legendre [22], Gauss [13], Abel [1] and Jacobi [16], which includes his lecture course of 1838.

It was shown by Landen (1775) that the length of arc of a hyperbola could be expressed as the difference of the lengths of two elliptic arcs. The change of variables involved is equivalent to that used by Lagrange (1784/5) in his application of the *AGM* algorithm. However, Lagrange used the transformation in much greater generality, and it was his idea that elliptic integrals could be calculated numerically by iterating the transformation. The connection with the result of Landen was made explicit by Legendre (1786).

By bringing together his own results and those of others the treatise of Legendre [22], and his earlier *Exercices de calcul integral* (1811/19), contributed substantially to the discoveries of Abel and Jacobi. The supplementary third volume of his treatise, published in 1828 when he was 76, contains the first account of their work in book form.

The most important contribution of Abel (1827) was not the replacement of elliptic integrals by elliptic functions, but the study of the latter in the complex domain. In this way he established their double periodicity, determined their zeros and poles and (besides much else) showed that they could be represented as quotients of infinite products.

The triple product formula of Jacobi (1829) identified these infinite products with infinite series, whose rapid convergence made them well suited for numerical computation. Infinite series of this type had in fact already appeared in the *Théorie analytique de la Chaleur* of Fourier (1822), and Proposition 3 had essentially been proved by Poisson (1827). Remarkable generalizations of the Jacobi triple product formula to affine Lie algebras have recently been obtained by Macdonald [23] and Kac and Peterson [17]. For an introductory account, see Neher [24].

It is difficult to understand the glee with which some authors attribute to Gauss results on elliptic functions, since the world owes its knowledge of these results not to him, but to others. Gauss's work was undoubtedly independent and in most cases earlier, although not in the case of the arithmetic-geometric mean. The remark, in §335 of his *Disquisitiones Arithmeticae* (1801), that his results on the division of the circle into n equal parts applied also to the lemniscate, was one of the motivations for Abel, who

carried out this extension. (For a modern account, see Rosen [25].) However, Gauss's claim in a letter to Schumacher of 30 May 1828, quoted in Krazer [20], that Abel had anticipated about a third of his own research is quite unjustified, and not only because of his inability to bring his work to a form in which it could be presented to the world.

It was *proved* by Liouville (1834) that elliptic integrals of the first and second kinds are always 'nonelementary'. For an introductory account of Liouville's theory, see Kasper [18]. (But elliptic integrals of the third kind may be 'elementary'; see Chapter IV, §7.)

The three kinds of elliptic integral may also be characterized function-theoretically. On the Riemann surface of the algebraic function $w^2 = g(z)$, where g is a cubic without repeated roots, the differential dz/w is everywhere holomorphic, the differential zdz/w is holomorphic except for a double pole at ∞ with zero residue, and the differential $[w(z)+w(a)]dz/2(z-a)w(z)$ is holomorphic except for two simple poles at a and ∞ with residues 1 and -1 respectively.

Many integrals which are not visibly elliptic may be reduced to elliptic integrals by a change of variables. A compilation is given by Byrd and Friedman [8], pp. 254–271.

The arithmetic-geometric mean may also be defined for pairs of complex numbers; a thorough discussion is given by Cox [9]. For the application of the *AGM* algorithm to integrals which are not strictly elliptic, see Bartky [4].

The differential equation (6) is a special case of the hypergeometric differential equation. In fact, if $|\lambda| < 1$, then by expanding $(1-\lambda x)^{-1/2}$, resp. $(1-\lambda x)^{1/2}$, by the binomial theorem and integrating term by term, the complete elliptic integrals

$$K(\lambda) = \int_0^1 [4x(1-x)(1-\lambda x)]^{-1/2}dx, \quad E(\lambda) = \int_0^1 [(1-\lambda x)/4x(1-x)]^{1/2}dx,$$

may be identified with the hypergeometric functions

$$(\pi/2)F(1/2, 1/2; 1; \lambda), \quad (\pi/2)F(-1/2, 1/2; 1; \lambda),$$

where

$$F(\alpha, \beta; \gamma; z) = 1 + \alpha\beta z/1\cdot\gamma + \alpha(\alpha+1)\beta(\beta+1)z^2/1\cdot 2\cdot\gamma(\gamma+1) + \cdots.$$

Many transformation formulas for the complete elliptic integrals may be regarded as special cases of more general transformation formulas for the hypergeometric function.

The proof in §3 that $K(1-\lambda)/K(\lambda)$ has positive real part is due to Falk [11].

It follows from (12)–(13) by induction that $S(nu)$ and $S'(nu)/S'(u)$ are rational functions of $S(u)$ for every integer n. The elliptic function $S(u)$ is said to admit *complex multiplication* if $S(\mu u)$ is a rational function of $S(u)$ for some complex number μ which is not an integer. It may be shown that $S(u)$ admits complex multiplication if and only if $\lambda \neq 0, 1$ and the period ratio $iK(1-\lambda)/K(\lambda)$ is a quadratic irrational, in the sense of Chapter IV. This condition is obviously satisfied if $\lambda = 1/2$, the case of the lemniscate.

A function $f(u)$ is said to possess an *algebraic addition theorem* if there is a polynomial $p(x, y, z)$, not identically zero and with coefficients independent of u, v, such that

$$p(f(u+v), f(u), f(v)) = 0 \quad \text{for all } u, v.$$

It may be shown that a function f, which is meromorphic in the whole complex plane, has an algebraic addition theorem if and only if it is either a rational function or, when the independent variable is scaled by a constant factor, a rational function of $S(u, \lambda)$ and its derivative $S'(u, \lambda)$ for some $\lambda \in \mathbb{C}$. This result (in different notation) is due to Weierstrass and is proved in Akhiezer [3], for example. A generalization of Weierstrass' theorem, due to Myrberg, is proved in Belavin and Drinfeld [6].

The term 'elliptic function' is often used to denote any function which is meromorphic in the whole complex plane and has two periods whose ratio is not real. It may be shown that, if the independent variable is scaled by a constant factor, an elliptic function in this general sense is a rational function of $S(u, \lambda)$ and $S'(u, \lambda)$ for some $\lambda \neq 0, 1$.

The functions $f(v)$ which are holomorphic in the whole complex plane $\mathbb{C}$ and satisfy the functional equations

$$f(v+1) = f(v), \quad f(v+\tau) = e^{-n\pi i(2v+\tau)} f(v),$$

where $n \in \mathbb{N}$ and $\tau \in \mathscr{H}$, form an n-dimensional complex vector space. It was shown by Hermite (1862) that this may be used to derive many relations between theta functions, such as Proposition 6.

Proposition 11 can be extended to give transformation formulas for the Jacobian functions when the parameter τ is multiplied by any positive integer n. See, for example, Tannery and Molk [27], vol. II.

The modular function was used by Picard (1879) to prove that a function $f(z)$, which is holomorphic for all $z \in \mathbb{C}$ and not a constant, assumes every complex value except perhaps one. The exponential function $\exp z$, which does not assume the value 0, illustrates that an exceptional value may exist. A careful proof of Picard's theorem is given in Ahlfors [2]. (There are also proofs which do not use the modular function.)

It was already observed by Lagrange (1813) that there is a correspondence between addition formulas for elliptic functions and the formulas of spherical trigonometry. This correspondence has been most intensively investigated by Study [26].

There is an n-dimensional generalization of theta functions, which has a useful application to the lattices studied in Chapter VIII. The theta function of an *integral lattice* Λ in $\mathbb{R}^n$ is defined by

$$\theta_\Lambda(\tau) = \sum_{u \in \Lambda} q^{(u,u)} = 1 + \sum_{m \geq 1} N_m q^m,$$

where $q = e^{\pi i \tau}$ and N_m is the number of vectors in Λ with square-norm m. If $n = 1$ and $\Lambda = \mathbb{Z}$, then

$$\theta_{\mathbb{Z}}(\tau) = 1 + 2q + 2q^4 + 2q^9 + \cdots = \theta(0; \tau).$$

It is easily seen that $\theta_\Lambda(\tau)$ is a holomorphic function of τ in the half-plane $\mathscr{I}\tau > 0$. It follows from Poisson's summation formula that the theta function of the *dual lattice* Λ^* is given by

$$\theta_{\Lambda^*}(\tau) = \mathrm{d}(\Lambda)(i/\tau)^{n/2} \theta_\Lambda(-1/\tau) \quad \text{for } \mathscr{I}\tau > 0.$$

Many geometrical properties of a lattice are reflected in its theta function. However, a lattice is not uniquely determined by its theta function, since there are lattices in $\mathbb{R}^4$ (and in higher dimensions) which are not isometric but have the same theta function.

For applications of elliptic functions and theta functions to classical mechanics, conformal mapping, geometry, theoretical chemistry, statistical mechanics and approximation theory, see Halphen [15] (vol. 2), Kober [19], Bos *et al.* [7], Glasser and Zucker [14], Baxter [5] and Todd [28]. Applications to number theory will be considered in the next chapter.

8 Selected References

[1] N.H. Abel, *Oeuvres complètes,* Tome 1, 2nd ed. (ed. L. Sylow et S. Lie), Grondahl, Christiania, 1881. [Reprinted J. Gabay, Sceaux, 1992]

[2] L.V. Ahlfors, *Complex analysis*, 3rd ed., McGraw-Hill, New York, 1979.

[3] N.I. Akhiezer, *Elements of the theory of elliptic functions*, American Mathematical Society, Providence, R.I., 1990. [English transl. of 2nd Russian edition, 1970]

[4] W. Bartky, Numerical calculation of a generalized complete elliptic integral, *Rev. Modern Phys*. **10** (1938), 264–269.

[5] R.J. Baxter, *Exactly solved models in statistical mechanics*, Academic Press, London, 1982. [Reprinted, 1989]

[6] A.A. Belavin and V.G. Drinfeld, Triangle equations and simple Lie algebras, *Soviet Sci. Rev. Sect. C: Math. Phys.* **4** (1984), 93–165. [Reprinted, Harwood, Amsterdam, 1998]

[7] H.J.M. Bos, C. Kers, F. Oort and D.W. Raven, Poncelet's closure theorem, *Exposition. Math.* **5** (1987), 289–364.

[8] P.F. Byrd and M.D. Friedman, *Handbook of elliptic integrals for engineers and scientists*, 2nd ed., Springer, Berlin, 1971.

[9] D.A. Cox, The arithmetic-geometric mean of Gauss, *Enseign. Math.* **30** (1984), 275–330.

[10] L. Euler, *Opera omnia*, Ser. I, Vol. XX (ed. A. Krazer), Leipzig, 1912.

[11] M. Falk, Beweis eines Satzes aus der Theorie der elliptischen Functionen, *Acta Math.* **7** (1885/6), 197–200.

[12] R. Fricke, *Elliptische Funktionen*, Encyklopädie der Mathematischen Wissenschaften, Band II, Teil 2, pp. 177–348, Teubner, Leipzig, 1921.

[13] C.F. Gauss, *Werke*, Band III, Göttingen, 1866. [Reprinted G. Olms, Hildesheim, 1973]

[14] M.L. Glasser and I.J. Zucker, Lattice sums, *Theoretical chemistry*: *Advances and perspectives* **5** (1980), 67–139.

[15] G.H. Halphen, *Traité des fonctions elliptiques et de leurs applications,* 3 vols., Gauthier-Villars, Paris, 1886–1891.

[16] C.G.J. Jacobi, *Gesammelte Werke,* Band I (ed. C.W. Borchardt), Berlin, 1881. [Reprinted Chelsea, New York, 1969]

[17] V.G. Kac and D.H. Peterson, Infinite-dimensional Lie algebras, theta functions and modular forms, *Adv. in Math*. **53** (1984), 125–264.

[18] T. Kasper, Integration in finite terms: the Liouville theory, *Math. Mag.* **53** (1980), 195–201.

[19] H. Kober, *Dictionary of conformal representations*, Dover, New York, 1952.

[20] A. Krazer, Zur Geschichte des Umkehrproblems der Integral, *Jahresber. Deutsch. Math.-Verein.* **18** (1909), 44–75.

[21] J.L. Lagrange, *Oeuvres,* t. 2 (ed. J.-A. Serret), Gauthier-Villars, Paris, 1868. [Reprinted G. Olms, Hildesheim, 1973]

[22] A.M. Legendre, *Traité des fonctions elliptiques et des intégrales Eulériennes, avec des tables pour en faciliter le calcul numérique*, Paris, t.1 (1825), t.2 (1826), t.3 (1828). [Microform, Readex Microprint Corporation, New York, 1970]

[23] I.G. Macdonald, Affine root systems and Dedekind's η-function, *Invent. Math.* **15** (1972), 91–143.
[24] E. Neher, Jacobis Tripelprodukt-Identität und η-Identitäten in der Theorie affiner Lie-Algebren, *Jahresber. Deutsch. Math.-Verein.* **87** (1985), 164–181.
[25] M. Rosen, Abel's theorem on the lemniscate, *Amer. Math. Monthly* **88** (1981), 387–395.
[26] E. Study, *Sphärische Trigonometrie, orthogonale Substitutionen und elliptische Funktionen*, Leipzig, 1893.
[27] J. Tannery and J. Molk, *Éléments de la théorie des fonctions elliptiques*, 4 vols., Gauthier-Villars, Paris, 1893–1902. [Reprinted Chelsea, New York, 1972]
[28] J. Todd, Applications of transformation theory: a legacy from Zolotarev (1847-1878), *Approximation theory and spline functions* (ed. S.P. Singh et al.), pp. 207–245, Reidel, Dordrecht, 1984.
[29] E.T. Whittaker and G.N. Watson, *A course of modern analysis*, 4th ed., Cambridge University Press, 1927. [Reprinted, 1996]

XIII

Connections with Number Theory

1 Sums of Squares

In Proposition II.40 we proved Lagrange's theorem that every positive integer can be represented as a sum of 4 squares. Jacobi (1829), at the end of his *Fundamenta Nova*, gave a completely different proof of this theorem with the aid of theta functions. Moreover, his proof provided a formula for the number of different representations. Hurwitz (1896), by developing further the arithmetic of quaternions which was used in Chapter II, also derived this formula. Here we give Jacobi's argument preference since, although it is less elementary, it is more powerful.

Proposition 1 *The number of representations of a positive integer m as a sum of* 4 *squares of integers is equal to* 8 *times the sum of those positive divisors of m which are not divisible by* 4.

Proof From the series expansion

$$\theta_{00}(0) = \sum_{n\in\mathbb{Z}} q^{n^2}$$

we obtain

$$\theta_{00}^4(0) = \sum_{n_1,\dots,n_4\in\mathbb{Z}} q^{n_1^2+\cdots+n_4^2} = 1 + \sum_{m\geq 1} r_4(m)q^m,$$

where $r_4(m)$ is the number of solutions in integers $n_1, \dots, n_4$ of the equation

$$n_1^2 + \cdots + n_4^2 = m.$$

We will prove the result by comparing this with another expression for $\theta_{00}^4(0)$.

We can write equation (43) of Chapter XII in the form

$$\theta_{10}'(v)/\theta_{10}(v) - \theta_{01}'(v)/\theta_{01}(v) = \pi i\theta_{00}^2(0)\theta_{00}(v)\theta_{11}(v)/\theta_{01}(v)\theta_{10}(v).$$

Differentiating with respect to v and then putting $v = 0$, we obtain

$$\theta_{10}''(0)/\theta_{10}(0) - \theta_{01}''(0)/\theta_{01}(0) = \pi i\theta_{00}^3(0)\theta_{11}'(0)/\theta_{01}(0)\theta_{10}(0) = -\pi^2\theta_{00}^4(0),$$

W.A. Coppel, *Number Theory: An Introduction to Mathematics*, Universitext,
DOI: 10.1007/978-0-387-89486-7_13, © Springer Science + Business Media, LLC 2009

by (36) of Chapter XII. Since the theta functions are all solutions of the partial differential equation

$$\partial^2 y/\partial v^2 = -4\pi^2 q \partial y/\partial q,$$

the last relation can be written in the form

$$4q\partial/\partial q \log\{\theta_{10}(0)/\theta_{01}(0)\} = \theta_{00}^4(0).$$

On the other hand, the product expansions of the theta functions show that

$$\begin{aligned}\theta_{10}(0)/\theta_{01}(0) &= 2q^{1/4}\prod_{n\geq 1}(1+q^{2n})^2 \Big/ \prod_{n\geq 1}(1-q^{2n-1})^2 \\ &= 2q^{1/4}\prod_{n\geq 1}(1-q^{4n})^2 \Big/ \prod_{n\geq 1}(1-q^{2n})^2(1-q^{2n-1})^2 \\ &= 2q^{1/4}\prod_{n\geq 1}(1-q^{4n})^2(1-q^n)^{-2}.\end{aligned}$$

Differentiating logarithmically, we obtain

$$\begin{aligned}\theta_{00}^4(0) &= 4q\partial/\partial q \log\{\theta_{10}(0)/\theta_{01}(0)\} \\ &= 1 + 8\sum_{n\geq 1} nq^n/(1-q^n) - 8\sum_{n\geq 1} 4nq^{4n}/(1-q^{4n}) \\ &= 1 + 8\sum_{n\geq 1}\sum_{k\geq 1}(nq^{kn} - 4nq^{4kn}) \\ &= 1 + 8\sum_{m\geq 1}\{\sigma(m) - \sigma'(m)\}q^m,\end{aligned}$$

where $\sigma(m)$ is the sum of all positive divisors of m and $\sigma'(m)$ is the sum of all positive divisors of m which are divisible by 4. Since the coefficients in a power series expansion are uniquely determined, it follows that

$$r_4(m) = 8\{\sigma(m) - \sigma'(m)\}. \qquad \square$$

Proposition 1 may also be restated in the form: the number of representations of m as a sum of 4 squares is equal to 8 times the sum of the odd positive divisors of m if m is odd, and 24 times this sum if m is even. For example,

$$r_4(10) = 24(1+5) = 144.$$

Since any positive integer has the odd positive divisor 1, Proposition 1 provides a new proof of Proposition II.40.

The number of representations of a positive integer as a sum of 2 squares may be treated in the same way, as Jacobi also showed (or, alternatively, by developing further the arithmetic of Gaussian integers):

Proposition 2 *The number of representations of a positive integer m as a sum of 2 squares of integers is equal to 4 times the excess of the number of positive divisors of m of the form $4h+1$ over the number of positive divisors of the form $4h+3$.*

Proof We have

$$\theta_{00}^2(0) = \sum_{n_1,n_2\in\mathbb{Z}} q^{n_1^2+n_2^2} = 1 + \sum_{m\geq 1} r_2(m)q^m,$$

where $r_2(m)$ is the number of solutions in integers n_1, n_2 of the equation

$$n_1^2 + n_2^2 = m.$$

To obtain another expression for $\theta_{00}^2(0)$ we use again the relation

$$\theta_{10}'(v)/\theta_{10}(v) - \theta_{01}'(v)/\theta_{01}(v) = \pi i\theta_{00}^2(0)\theta_{00}(v)\theta_{11}(v)/\theta_{01}(v)\theta_{10}(v),$$

but this time we simply take $v = 1/4$. Since

$$\theta_{01}(1/4) = \sum_{n\in\mathbb{Z}}(-i)^n q^{n^2} = \sum_{n\in\mathbb{Z}} i^{-n}q^{n^2} = \theta_{00}(1/4),$$

and similarly $\theta_{11}(1/4) = i\ \theta_{10}(1/4)$, we obtain

$$\pi\theta_{00}^2(0) = \theta_{01}'(1/4)/\theta_{01}(1/4) - \theta_{10}'(1/4)/\theta_{10}(1/4).$$

By differentiating logarithmically the product expansion for $\theta_{10}(v)$ and then putting $v = 1/4$, we get

$$\theta_{10}'(1/4)/\theta_{10}(1/4) = -\pi - 4\pi\sum_{n\geq 1} q^{2n}/(1+q^{4n}).$$

Similarly, by differentiating logarithmically the product expansion for $\theta_{01}(v)$ and then putting $v = 1/4$, we get

$$\theta_{01}'(1/4)/\theta_{01}(1/4) = 4\pi\sum_{n\geq 1} q^{2n-1}/(1+q^{4n-2}).$$

Thus

$$\theta_{01}'(1/4)/\theta_{01}(1/4) - \theta_{10}'(1/4)/\theta_{10}(1/4) = \pi + 4\pi\sum_{n\geq 1} q^n/(1+q^{2n})$$

and hence

$$\theta_{00}^2(0) = 1 + 4\sum_{n\geq 1} q^n/(1+q^{2n}).$$

Since

$$q^n/(1+q^{2n}) = q^n(1-q^{2n})/(1-q^{4n}) = (q^n - q^{3n})\sum_{k\geq 0} q^{4kn},$$

it follows that

$$\theta_{00}^2(0) = 1 + 4\sum_{n\geq 1}\sum_{k\geq 0}\{q^{(4k+1)n} - q^{(4k+3)n}\}$$
$$= 1 + 4\sum_{m\geq 1}\{d_1(m) - d_3(m)\}q^m,$$

where $d_1(m)$ and $d_3(m)$ are respectively the number of positive divisors of m congruent to 1 and 3 mod 4. Hence

$$r_2(m) = 4\{d_1(m) - d_3(m)\}. \qquad \square$$

From Proposition 2 we immediately obtain again that any prime $p \equiv 1 \bmod 4$ may be represented as a sum of 2 squares and that the representation is essentially unique. Proposition II.39 may also be rederived.

The number $r_s(m)$ of representations of a positive integer m as a sum of s squares has been expressed by explicit formulas for many other values of s besides 2 and 4. Systematic ways of attacking the problem are provided by the theory of modular forms and the circle method of Hardy, Ramanujan and Littlewood.

2 Partitions

A *partition* of a positive integer n is a set of positive integers with sum n. For example, $\{2, 1, 1\}$ is a partition of 4. We denote the number of distinct partitions of n by $p(n)$. For example, $p(4) = 5$, since all partitions of 4 are given by

$$\{4\}, \{3, 1\}, \{2, 2\}, \{2, 1, 1\}, \{1, 1, 1, 1\}.$$

It was shown by Euler (1748) that the sequence $p(n)$ has a simple *generating function*:

Proposition 3 *If* $|x| < 1$, *then*

$$1/(1-x)(1-x^2)(1-x^3)\cdots = 1 + \sum_{n\geq 1} p(n)x^n.$$

Proof If $|x| < 1$, then the infinite product $\prod_{m\geq 1}(1-x^m)$ converges and its reciprocal has a convergent power series expansion. To determine the coefficients of this expansion note that, since

$$(1-x^m)^{-1} = \sum_{k\geq 0} x^{km},$$

the coefficient of $x^n (n \geq 1)$ in the product $\prod_{m\geq 1}(1-x^m)^{-1}$ is the number of representations of n in the form

$$n = 1k_1 + 2k_2 + \cdots,$$

where the k_j are non-negative integers. But this number is precisely $p(n)$, since any partition is determined by the number of 1's, 2's, ... that it contains. $\square$

For many purposes the discussion of convergence is superfluous and Proposition 3 may be regarded simply as a relation between formal products and formal power series.

Euler also obtained an interesting counterpart to Proposition 3, which we will derive from Jacobi's triple product formula.

Proposition 4 *If* $|x| < 1$, *then*

$$(1-x)(1-x^2)(1-x^3)\cdots = \sum_{m\in\mathbb{Z}} (-1)^m x^{m(3m+1)/2}.$$

Proof If we take $q = x^{3/2}$ and $z = -x^{1/2}$ in Proposition XII.2, we obtain at once the result, since

$$\prod_{n\geq 1}(1-x^{3n})(1-x^{3n-1})(1-x^{3n-2}) = \prod_{k\geq 1}(1-x^k). \qquad \square$$

Proposition 4 also has a combinatorial interpretation. The coefficient of $x^n (n \geq 1)$ in the power series expansion of $\prod_{k\geq 1}(1-x^k)$ is

$$s_n = \sum (-1)^v,$$

where the sum is over all partitions of n into *unequal* parts and v is the number of parts in the partition. In other words,

$$s_n = p_e^*(n) - p_o^*(n),$$

where $p_e^*(n)$, resp. $p_o^*(n)$, is the number of partitions of the positive integer n into an even, resp. odd, number of unequal parts. On the other hand,

$$\sum_{m\in\mathbb{Z}} (-1)^m x^{m(3m+1)/2} = 1 + \sum_{m\geq 1} (-1)^m \{x^{m(3m+1)/2} + x^{m(3m-1)/2}\}.$$

Thus Proposition 4 says that $p_e^*(n) = p_o^*(n)$ unless $n = m(3m \pm 1)/2$ for some $m \in \mathbb{N}$, in which case $p_e^*(n) - p_o^*(n) = (-1)^m$.

From Propositions 3 and 4 we obtain

$$\left[1 + \sum_{m\geq 1} (-1)^m \{x^{m(3m+1)/2} + x^{m(3m-1)/2}\}\right]\left[1 + \sum_{k\geq 1} p(k)x^k\right] = 1.$$

Multiplying out on the left side and equating to zero the coefficient of $x^n (n \geq 1)$, we obtain the recurrence relation:

$$\begin{aligned} p(n) = {} & p(n-1) + p(n-2) - p(n-5) - p(n-7) \\ & + \cdots + (-1)^{m-1} p(n - m(3m-1)/2) \\ & + (-1)^{m-1} p(n - m(3m+1)/2) + \cdots, \end{aligned}$$

where $p(0) = 1$ and $p(k) = 0$ for $k < 0$. This recurrence relation is quite an efficient way of calculating $p(n)$. It was used by MacMahon (1918) to calculate $p(n)$ for $n \leq 200$.

In the same way that we proved Proposition 3 we may show that, if $|x| < 1$, then

$$1/(1-x)(1-x^2)\cdots(1-x^m) = 1 + \sum_{n\geq 1} p_m(n)x^n,$$

where $p_m(n)$ is the number of partitions of n into parts not exceeding m.

From the vast number of formulas involving partitions and their generating functions we select only one more pair, the celebrated *Rogers–Ramanujan identities*. The proof of these identities will be based on the following preliminary result:

Proposition 5 *If $|q| < 1$ and $|x| < |q|^{-1}$, then*

$$1 + \sum_{n\geq 1} x^n q^{n^2}/(q)_n = \sum_{n\geq 0} (-1)^n x^{2n} q^{5n(n+1)/2-2n}\{1 - x^2 q^{2(2n+1)}\}/(q)_n (xq^{n+1})_\infty,$$

where $(a)_0 = 1$,

$$(a)_n = (1-a)(1-aq)\cdots(1-aq^{n-1}) \quad \text{if } n \geq 1, \text{ and}$$
$$(a)_\infty = (1-a)(1-aq)(1-aq^2)\cdots.$$

Proof Consider the *q-difference equation*

$$f(x) = f(xq) + xq\ f(xq^2).$$

A formal power series $\sum_{n\geq 0} a_n x^n$ satisfies this equation if and only if

$$a_n(1-q^n) = a_{n-1}q^{2n-1} \quad (n \geq 1).$$

Thus the only formal power series solution with $a_0 = 1$ is

$$f(x) = 1 + xq/(1-q) + x^2q^4/(1-q)(1-q^2) + x^3q^9/(1-q)(1-q^2)(1-q^3) + \cdots.$$

Moreover, if $|q| < 1$, this power series converges for all $x \in \mathbb{C}$.

If $|q| < 1$, the functions

$$F(x) = \sum_{n\geq 0} (-1)^n x^{2n} q^{5n(n+1)/2-2n}\{1 - x^2 q^{2(2n+1)}\}/(q)_n (xq^{n+1})_\infty,$$
$$G(x) = \sum_{n\geq 0} (-1)^n x^{2n} q^{5n(n+1)/2-n}\{1 - xq^{2n+1}\}/(q)_n (xq^{n+1})_\infty$$

are holomorphic for $|x| < |q|^{-1}$.

We have

$$\begin{aligned}
&F(x)-G(x)\\
&\quad=\sum_{n\geq 0}(-1)^n x^{2n}q^{5n(n+1)/2}\{q^{-2n}-x^2q^{2(n+1)}-q^{-n}+xq^{n+1}\}/(q)_n(xq^{n+1})_\infty\\
&\quad=\sum_{n\geq 0}(-1)^n x^{2n}q^{5n(n+1)/2}\{q^{-2n}(1-q^n)+xq^{n+1}(1-xq^{n+1})\}/(q)_n(xq^{n+1})_\infty\\
&\quad=\sum_{n\geq 1}(-1)^n x^{2n}q^{5n(n+1)/2-2n}/(q)_{n-1}(xq^{n+1})_\infty\\
&\qquad+xq\sum_{n\geq 0}(-1)^n x^{2n}q^{5n(n+1)/2+n}/(q)_n(xq^{n+2})_\infty\\
&\quad=-x^2q^3\sum_{n\geq 0}(-1)^n x^{2n}q^{5n(n+1)/2+3n}/(q)_n(xq^{n+2})_\infty\\
&\qquad+xq\sum_{n\geq 0}(-1)^n x^{2n}q^{5n(n+1)/2+n}/(q)_n(xq^{n+2})_\infty\\
&\quad=xq\sum_{n\geq 0}(-1)^n (xq)^{2n}q^{5n(n+1)/2-n}\{1-(xq)q^{2n+1}\}/(q)_n(xq^{n+2})_\infty\\
&\quad=xqG(xq).
\end{aligned}$$

Similarly,

$$\begin{aligned}
G(x)&=\sum_{n\geq 0}(-1)^n x^{2n}q^{5n(n+1)/2}\{q^{-n}-xq^{n+1}\}/(q)_n(xq^{n+1})_\infty\\
&=\sum_{n\geq 0}(-1)^n x^{2n}q^{5n(n+1)/2}\{q^{-n}(1-q^n)+1-xq^{n+1}\}/(q)_n(xq^{n+1})_\infty\\
&=\sum_{n\geq 1}(-1)^n x^{2n}q^{5n(n+1)/2-n}/(q)_{n-1}(xq^{n+1})_\infty\\
&\quad+\sum_{n\geq 0}(-1)^n x^{2n}q^{5n(n+1)/2}/(q)_n(xq^{n+2})_\infty\\
&=\sum_{n\geq 0}(-1)^n (xq)^{2n}q^{5n(n+1)/2-2n}\{1-(xq)^2q^{2(2n+1)}\}/(q)_n(xq^{n+2})_\infty\\
&=F(xq).
\end{aligned}$$

Combining this with the previous relation, we obtain

$$F(x)=F(xq)+xq\ F(xq^2).$$

But we have seen that this q-difference equation has a unique holomorphic solution $f(x)$ such that $f(0)=1$. Hence $F(x)=f(x)$. □

The Rogers–Ramanujan identities may now be easily derived:

Proposition 6 *If* $|q| < 1$, *then*

$$\sum_{n\geq 0} q^{n^2}/(1-q)(1-q^2)\cdots(1-q^n) = \prod_{m\geq 0}(1-q^{5m+1})^{-1}(1-q^{5m+4})^{-1},$$
$$\sum_{n\geq 0} q^{n(n+1)}/(1-q)(1-q^2)\cdots(1-q^n) = \prod_{m\geq 0}(1-q^{5m+2})^{-1}(1-q^{5m+3})^{-1}.$$

Proof Put $P = \prod_{k\geq 1}(1-q^k)$. By Proposition 5 and its proof we have

$$\sum_{n\geq 0} q^{n^2}/(1-q)(1-q^2)\cdots(1-q^n) = F(1)$$
$$= \left[1 + \sum_{n\geq 1}(-1)^n\{q^{n(5n+1)/2} + q^{n(5n-1)/2}\}\right]\Big/ P$$

and, since $F(q) = G(1)$,

$$\sum_{n\geq 0} q^{n(n+1)}/(1-q)(1-q^2)\cdots(1-q^n) = F(q)$$
$$= \left[1 + \sum_{n\geq 1}(-1)^n\{q^{n(5n+3)/2} + q^{n(5n-3)/2}\}\right]\Big/ P.$$

On the other hand, by replacing q by $q^{5/2}$ and z by $-q^{1/2}$, resp. $-q^{3/2}$, in Jacobi's triple product formula (Proposition XII.2), we obtain

$$\sum_{n\in\mathbb{Z}}(-1)^n q^{n(5n+1)/2} = \prod_{m\geq 1}(1-q^{5m})(1-q^{5m-2})(1-q^{5m-3})$$
$$= P\Big/\prod_{m\geq 0}(1-q^{5m+1})(1-q^{5m+4})$$

and

$$\sum_{n\in\mathbb{Z}}(-1)^n q^{n(5n+3)/2} = \prod_{m\geq 1}(1-q^{5m})(1-q^{5m-1})(1-q^{5m-4})$$
$$= P\Big/\prod_{m\geq 0}(1-q^{5m+2})(1-q^{5m+3}).$$

Combining these relations with the previous ones, we obtain the result. □

The combinatorial interpretation of the Rogers–Ramanujan identities was pointed out by MacMahon (1916). The first identity says that the number of partitions of a positive integer n into parts congruent to $\pm 1 \bmod 5$ is equal to the number of partitions of n into parts that differ by at least 2. The second identity says that the number of partitions of a positive integer n into parts congruent to $\pm 2 \bmod 5$ is equal to the number of partitions of n into parts greater than 1 that differ by at least 2.

A remarkable application of the Rogers–Ramanujan identities to the hard hexagon model of statistical mechanics was found by Baxter (1981). Many other models in statistical mechanics have been exactly solved with the aid of theta functions. A unifying principle is provided by the vast theory of infinite-dimensional Lie algebras which has been developed over the past 25 years.

The number $p(n)$ of partitions of n increases rapidly with n. It was first shown by Hardy and Ramanujan (1918) that

$$p(n) \sim e^{\pi\sqrt{2n/3}}/4n\sqrt{3} \quad \text{as } n \to \infty.$$

They further obtained an asymptotic series for $p(n)$, which was modified by Rademacher (1937) into a convergent series, from which it is even possible to calculate $p(n)$ exactly. A key role in the difficult proof is played by the behaviour under transformations of the modular group of *Dedekind's eta function*

$$\eta(\tau) = q^{1/12}\prod_{k\geq 1}(1-q^{2k}),$$

where $q = e^{\pi i\tau}$ and $\tau \in \mathscr{H}$ (the upper half-plane).

The paper of Hardy and Ramanujan contained the first use of the 'circle method', which was subsequently applied by Hardy and Littlewood to a variety of problems in analytic number theory.

3 Cubic Curves

We define an *affine plane curve* over a field K to be a polynomial $f(X, Y)$ in two indeterminates with coefficients from K, but we regard two polynomials $f(X, Y)$ and $f^*(X, Y)$ as defining the same affine curve if $f^* = \lambda f$ for some nonzero $\lambda \in K$. The *degree* of the curve is defined without ambiguity to be the degree of the polynomial f.

If

$$f(X, Y) = aX + bY + c$$

is a polynomial of degree 1, the curve is said to be an *affine line*. If

$$f(X, Y) = aX^2 + bXY + cY^2 + lX + mY + n$$

is a polynomial of degree 2, the curve is said to be an *affine conic*. If $f(X, Y)$ is a polynomial of degree 3, the curve is said to be an *affine cubic*. It is the cubic case in which we will be most interested.

Let $\mathscr{C}$ be an affine plane curve over the field K, defined by the polynomial $f(X, Y)$. We say that $(x, y) \in K^2$ is a *point* or, more precisely, a *K-point* of the affine curve $\mathscr{C}$ if $f(x, y) = 0$. The K-point (x, y) is said to be *non-singular* if there exist $a, b \in K$, not both zero, such that

$$f(x + X, y + Y) = aX + bY + \cdots,$$

where all unwritten terms have degree > 1. Since a, b are uniquely determined by f, we can define the *tangent* to the affine curve $\mathscr{C}$ at the non-singular point (x, y) to be the affine line

$$\ell(X, Y) = aX + bY - (ax + by).$$

It is easily seen that these definitions do not depend on the choice of polynomial within an equivalence class $\{\lambda f : 0 \neq \lambda \in K\}$.

The study of the asymptotes of an affine plane curve leads one to consider also its 'points at infinity', the asymptotes being the tangents at these points. We will now make this precise.

If the polynomial $f(X, Y)$ has degree d, then

$$F(X, Y, Z) = Z^d f(X/Z, Y/Z)$$

is a homogeneous polynomial of degree d such that

$$f(X, Y) = F(X, Y, 1).$$

Furthermore, if $\mathscr{F}(X, Y, Z)$ is any homogeneous polynomial such that $f(X, Y) = \mathscr{F}(X, Y, 1)$, then $\mathscr{F}(X, Y, Z) = Z^m F(X, Y, Z)$ for some non-negative integer m.

We define a *projective plane curve* over a field K to be a homogeneous polynomial $F(X, Y, Z)$ of degree $d > 0$ in three indeterminates with coefficients from K, but we regard two homogeneous polynomials $F(X, Y, Z)$ and $F^*(X, Y, Z)$ as defining the same projective curve if $F^* = \lambda F$ for some nonzero $\lambda \in K$. The projective curve is said to be a *projective line, conic* or *cubic* if F has degree 1,2 or 3 respectively.

If $\mathscr{C}$ is an affine plane curve, defined by a polynomial $f(X, Y)$ of degree $d > 0$, the projective plane curve $\bar{\mathscr{C}}$, defined by the homogeneous polynomial $Z^d f(X/Z, Y/Z)$ of the same degree, is called the *projective completion* of $\mathscr{C}$. Thus the projective completion of an affine line, conic or cubic is respectively a projective line, conic or cubic.

Let $\bar{\mathscr{C}}$ be a projective plane curve over the field K, defined by the homogeneous polynomial $F(X, Y, Z)$. We say that $(x, y, z) \in K^3$ is a *point*, or *K-point*, of $\bar{\mathscr{C}}$ if $(x, y, z) \neq (0, 0, 0)$ and $F(x, y, z) = 0$, but we regard two triples (x, y, z) and (x^*, y^*, z^*) as defining the same K-point if

$$x^* = \lambda x, y^* = \lambda y, z^* = \lambda z \quad \text{for some nonzero } \lambda \in K.$$

If $\bar{\mathscr{C}}$ is the projective completion of the affine plane curve $\mathscr{C}$, then a point (x, y, z) of $\bar{\mathscr{C}}$ with $z \neq 0$ corresponds to a point $(x/z, y/z)$ of $\mathscr{C}$, and a point $(x, y, 0)$ of $\bar{\mathscr{C}}$ corresponds to a *point at infinity* of $\mathscr{C}$.

The K-point (x, y, z) of the projective plane curve defined by the homogeneous polynomial $F(X, Y, Z)$ is said to be *non-singular* if there exist $a, b, c \in K$, not all zero, such that

$$F(x + X, y + Y, z + Z) = aX + bY + cZ + \cdots,$$

where all unwritten terms have degree > 1. Since a, b, c are uniquely determined by F, we can define the *tangent* to the projective curve at the non-singular point (x, y, z)

to be the projective line defined by $aX + bY + cZ$. It follows from Euler's theorem on homogeneous functions that (x, y, z) is itself a point of the tangent.

It is easily seen that if $\bar{\mathscr{C}}$ is the projective completion of an affine plane curve $\mathscr{C}$, and if $z \neq 0$, then (x, y, z) is a non-singular point of $\bar{\mathscr{C}}$ if and only if $(x/z, y/z)$ is a non-singular point of $\mathscr{C}$. Moreover, if the tangent to $\bar{\mathscr{C}}$ at (x, y, z) is the projective line

$$\bar{\ell}(X, Y, Z) = aX + bY + cZ,$$

then the tangent to $\mathscr{C}$ at $(x/z, y/z)$ is the affine line defined by

$$\ell(X, Y) = aX + bY + c.$$

Let $\mathscr{C}$ be an affine plane curve over the field K, defined by the polynomial $f(X, Y)$, and let (x, y) be a non-singular K-point of $\mathscr{C}$. Then we can write

$$f(x + X, y + Y) = aX + bY + f_2(X, Y) + \cdots,$$

where a, b are not both zero, $f_2(X, Y)$ is a homogeneous polynomial of degree 2, and all unwritten terms have degree > 2. The non-singular point (x, y) is said to be an *inflection point* or, more simply, a *flex* of $\mathscr{C}$ if $f_2(X, Y)$ is divisible by $aX + bY$.

Similarly we can define a flex for a projective plane curve. Let (x, y, z) be a non-singular point of the projective plane curve over the field K, defined by the homogeneous polynomial $F(X, Y, Z)$. Then we can write

$$F(x + X, y + Y, z + Z) = aX + bY + cZ + F_2(X, Y, Z) + \cdots,$$

where a, b, c are not all zero, $F_2(X, Y, Z)$ is a homogeneous polynomial of degree 2, and all unwritten terms have degree > 2. The non-singular point (x, y, z) is said to be a *flex* if $F_2(X, Y, Z)$ is divisible by $aX + bY + cZ$.

Two more definitions are required before we embark on our study of cubic curves. A projective curve over the field K, defined by the homogeneous polynomial $F(X, Y, Z)$ of degree $d > 0$, is said to be reducible over K if

$$F(X, Y, Z) = F_1(X, Y, Z)F_2(X, Y, Z),$$

where F_1 and F_2 are homogeneous polynomials of degree less than d with coefficients from K. The K-points of the curve defined by F are then just the K-points of the curve defined by F_1, together with the K-points of the curve defined by F_2. A curve is said to be *irreducible over* K if it is not reducible over K.

Two projective curves over the field K, defined by the homogeneous polynomials $F(X, Y, Z)$ and $G(X', Y', Z')$, are said to be *projectively equivalent* if there exists an invertible linear transformation

$$\begin{aligned} X &= a_{11}X' + a_{12}Y' + a_{13}Z' \\ Y &= a_{21}X' + a_{22}Y' + a_{23}Z' \\ Z &= a_{31}X' + a_{32}Y' + a_{33}Z' \end{aligned}$$

with coefficients $a_{ij} \in K$ such that

$$F(a_{11}X' + \cdots, a_{21}X' + \cdots, a_{31}X' + \cdots) = G(X', Y', Z').$$

It is clear that F and G necessarily have the same degree, and that projective equivalence is in fact an equivalence relation.

Consider now the affine cubic curve $\mathscr{C}$ defined by the polynomial

$$f(X, Y) = a_{30}X^3 + a_{21}X^2Y + a_{12}XY^2 + a_{03}Y^3 + a_{20}X^2 + a_{11}XY + a_{02}Y^2 + a_{10}X + a_{01}Y + a_{00}.$$

We assume that $\mathscr{C}$ has a non-singular K-point which is a flex. Without loss of generality, suppose that this is the origin. Then $a_{00} = 0$, a_{10} and a_{01} are not both zero, and

$$a_{20}X^2 + a_{11}XY + a_{02}Y^2 = (a_{10}X + a_{01}Y)(a'_{10}X + a'_{01}Y)$$

for some $a'_{10}, a'_{01} \in K$. By an invertible linear change of variables we may suppose that $a_{10} = 0, a_{01} = 1$. Then f has the form

$$f(X, Y) = Y + a_1XY + a_3Y^2 - a_0X^3 - a_2X^2Y - a_4XY^2 - a_6Y^3.$$

If $a_0 = 0$, then f is divisible by Y and the corresponding projective curve is reducible. Thus we now assume $a_0 \neq 0$. In fact we may assume $a_0 = 1$, by replacing f by a constant multiple and then scaling Y. The projective completion $\bar{\mathscr{C}}$ of $\mathscr{C}$ is now defined by the homogeneous polynomial

$$YZ^2 + a_1XYZ + a_3Y^2Z - X^3 - a_2X^2Y - a_4XY^2 - a_6Y^3.$$

If we interchange Y and Z, the flex becomes the unique point at infinity of the affine cubic curve defined by the polynomial

$$Y^2 + a_1XY + a_3Y - (X^3 + a_2X^2 + a_4X + a_6).$$

This can be further simplified by making mild restrictions on the field K. If K has characteristic $\neq 2$, i.e. if $1 + 1 \neq 0$, then by replacing Y by $(Y - a_1X - a_3)/2$ we obtain the cubic curve defined by the polynomial

$$Y^2 - (4X^3 + b_2X^2 + 2b_4X + b_6).$$

If K also has characteristic $\neq 3$, i.e. if $1 + 1 + 1 \neq 0$, then by replacing X by $(X - 3b_2)/6^2$ and Y by $2Y/6^3$, we obtain the cubic curve defined by the polynomial $Y^2 - (X^3 + aX + b)$. Thus we have proved:

Proposition 7 *If a projective cubic curve over the field K is irreducible and has a non-singular K-point which is a flex, then it is projectively equivalent to the projective completion $\mathscr{W} = \mathscr{W}(a_1, \ldots, a_6)$ of an affine curve of the form*

$$Y^2 + a_1XY + a_3Y - (X^3 + a_2X^2 + a_4X + a_6).$$

If K has characteristic $\neq 2, 3$, then it is projectively equivalent to the projective completion $\mathscr{C} = \mathscr{C}_{a,b}$ of an affine curve of the form

$$Y^2 - (X^3 + aX + b).$$

It is easily seen that, conversely, for any choice of $a_1, \ldots, a_6 \in K$ the curve $\mathscr{W}$, and in particular $\mathscr{C}_{a,b}$, is irreducible over K and that $\boldsymbol{0}$, the unique point at infinity, is a flex.

For any $u, r, s, t \in K$ with $u \neq 0$, the invertible linear change of variables

$$\begin{aligned} X &= u^2X' + r, \\ Y &= u^3Y' + su^2X' + t \end{aligned}$$

replaces the curve $\mathscr{W} = \mathscr{W}(a_1, \ldots, a_6)$ by a curve $\mathscr{W}' = \mathscr{W}'(a'_1, \ldots, a'_6)$ of the same form. The numbering of the coefficients reflects the fact that if $r = s = t = 0$, then

$$\begin{aligned} &a_1 = ua'_1, \quad a_2 = u^2a'_2, \quad a_3 = u^3a'_3, \\ &a_4 = u^4a'_4, \quad a_6 = u^6a'_6. \end{aligned}$$

In particular, for any nonzero $u \in K$, the invertible linear change of variables

$$\begin{aligned} X &= u^2X', \\ Y &= u^3Y' \end{aligned}$$

replaces $\mathscr{C}_{a,b}$ by $\mathscr{C}_{a',b'}$, where

$$\begin{aligned} a &= u^4a', \\ b &= u^6b'. \end{aligned}$$

By replacing X by $x + X$ and Y by $y + Y$, we see that if a K-point (x, y) of $\mathscr{C}_{a,b}$ is singular, then

$$3x^2 + a = y = 0,$$

which implies $4a^3 + 27b^2 = 0$. Thus the curve $\mathscr{C}_{a,b}$ has no singular points if $4a^3 + 27b^2 \neq 0$.

We will call

$$d := 4a^3 + 27b^2$$

the *discriminant* of the curve $\mathscr{C}_{a,b}$. It is not difficult to verify that if the cubic polynomial $X^3 + aX + b$ has roots e_1, e_2, e_3, then

$$d = -[(e_1 - e_2)(e_1 - e_3)(e_2 - e_3)]^2.$$

If $d = 0$, $a \neq 0$, then the polynomial $X^3 + aX + b$ has the repeated root $x_0 = -3b/2a$ and $P = (x_0, 0)$ is the unique singular point. If $d = a = 0$, then $b = 0$ and $P = (0, 0)$ is the unique singular point.

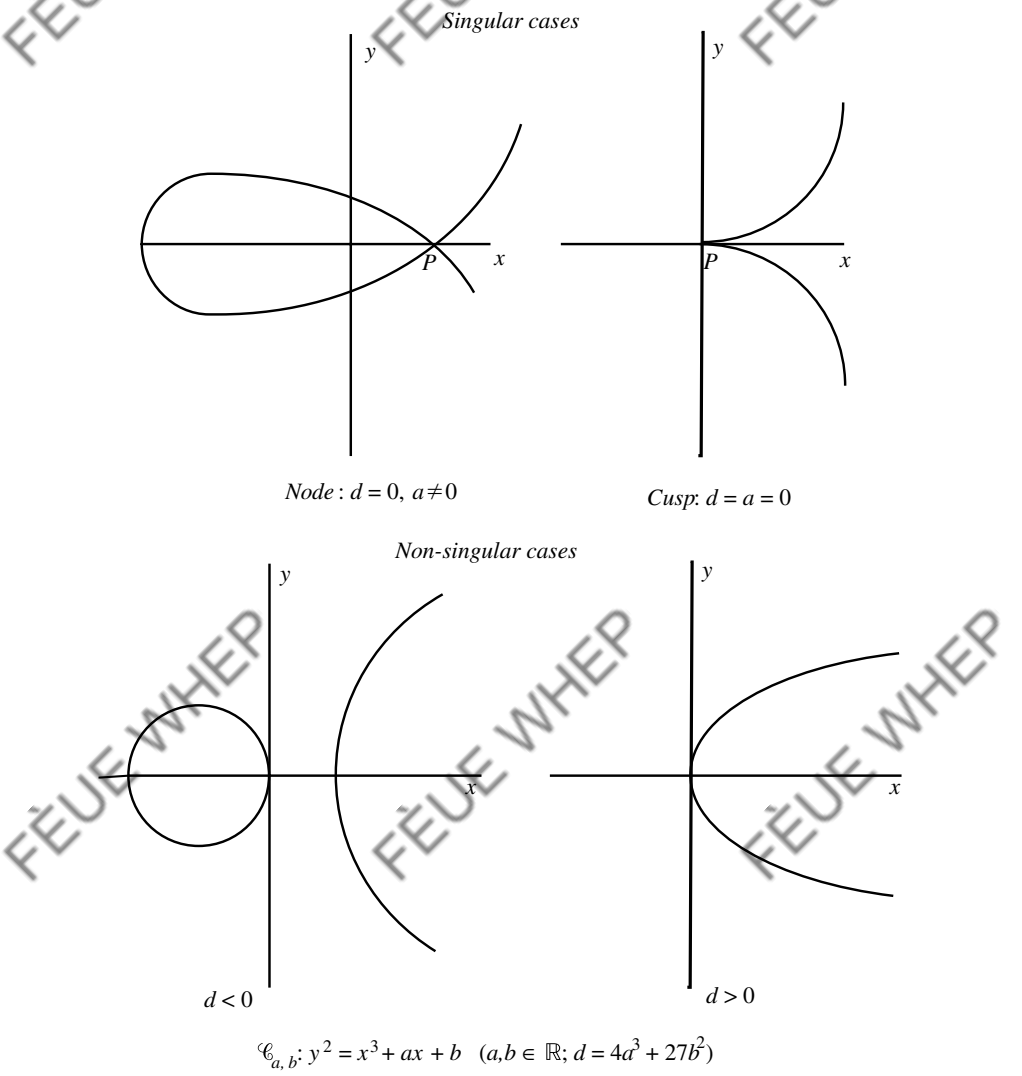

Fig. 1. Cubic curves over $\mathbb{R}$.

The different types of curve which arise when $K = \mathbb{R}$ is the field of real numbers are illustrated in Figure 1. The unique point at infinity $\boldsymbol{0}$ may be thought of as being at both ends of the y-axis. (In the case of a node, Figure 1 illustrates the situation for $x_0 > 0$. For $x_0 < 0$ the singular point is an isolated point of the curve.)

Suppose now that K is any field of characteristic $\neq 2, 3$ and that the curve $\mathscr{C}_{a,b}$ has zero discriminant. Because of the geometrical interpretation when $K = \mathbb{R}$, the unique singular point of the curve $\mathscr{C}_{a,b}$ is said to be a *node* if $a \neq 0$ and a *cusp* if $a = 0$. In the cusp case, if we put $T = Y/X$, then the cubic curve has the parametrization

$X = T^2$, $Y = T^3$. In the node case, if we put $T = Y/(X + 3b/2a)$, then it has the parametrization

$$X = T^2 + 3b/a, Y = T^3 + 9bT/2a.$$

Thus in both cases the cubic curve is in fact elementary.

We now restrict attention to non-singular cubic curves, i.e. curves which do not have a singular point.

Two K-points of a projective cubic curve determine a projective line, which intersects the curve in a third K-point. This procedure for generating additional K-points was used implicitly by Diophantus and explicitly by Newton. There is also another procedure, which may be regarded as a limiting case: the tangent to a projective cubic curve at a K-point intersects the curve in another K-point. The combination of the two procedures is known as the 'chord and tangent' process. It will now be described analytically for the cubic curve $\mathscr{C}_{a,b}$.

If $\boldsymbol{O}$ is the unique point at infinity of the cubic curve $\mathscr{C}_{a,b}$ and if $P = (x, y)$ is any finite K-point, then the affine line determined by $\boldsymbol{O}$ and P is $X - x$ and its other point of intersection with $\mathscr{C}_{a,b}$ is $P^* = (x, -y)$.

Now let $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ be any two finite K-points. If $x_1 \neq x_2$, then the affine line determined by P_1 and P_2 is

$$Y - mX - c,$$

where

$$m = (y_2 - y_1)/(x_2 - x_1), \quad c = (y_1x_2 - y_2x_1)/(x_2 - x_1),$$

and its third point of intersection with $\mathscr{C}_{a,b}$ is $P_3 = (x_3, y_3)$, where

$$x_3 = m^2 - x_1 - x_2, \quad y_3 = mx_3 + c.$$

If $x_1 = x_2$, but $y_1 \neq y_2$, then the affine line determined by P_1 and P_2 is $X - x_1$ and its other point of intersection with $\mathscr{C}_{a,b}$ is $\boldsymbol{O}$. Finally, if $P_1 = P_2$, it may be verified that the tangent to $\mathscr{C}_{a,b}$ at P_1 is the affine line

$$Y - mX - c,$$

where

$$m = (3x_1^2 + a)/2y_1, \quad c = (-x_1^3 + ax_1 + 2b)/2y_1,$$

and its other point of intersection with $\mathscr{C}_{a,b}$ is the point $P_3 = (x_3, y_3)$, where x_3 and y_3 are given by the same formulas as before, but with the new values of m and c (and with $x_2 = x_1$).

It is rather remarkable that the K-points of a non-singular projective cubic curve can be given the structure of an abelian group. That this is possible is suggested by the addition theorem for elliptic functions.

Suppose that $K = \mathbb{C}$ is the field of complex numbers and that the cubic curve is the projective completion $\mathscr{C}_\lambda$ of the affine curve

$$Y^2 - g_\lambda(X),$$

where

$$g_\lambda(X) = 4\lambda X^3 - 4(1+\lambda)X^2 + 4X$$

is Riemann's normal form and $\lambda \neq 0, 1$. If $S(u)$ is the elliptic function defined in §3 of Chapter XII, then $P(u) = (S(u), S'(u))$ is a point of $\mathscr{C}_\lambda$ for any $u \in \mathbb{C}$. If we define the sum of $P(u)$ and $P(v)$ to be the point $P(u+v)$, then the set of all $\mathbb{C}$-points of $\mathscr{C}_\lambda$ becomes an abelian group, with $P(0) = (0, 0)$ as identity element and with $P(-u) = (S(u), -S'(u))$ as the inverse of $P(u)$. In order to carry this construction over to the cubic curve $\mathscr{C}_{a,b}$ and to other fields than $\mathbb{C}$, we interpret it geometrically.

It was shown in (10) of Chapter XII that

$$S(u+v) = 4S(u)S(v)[S(v) - S(u)]^2/[S'(u)S(v) - S'(v)S(u)]^2.$$

The points $(x_1, y_1) = (S(u), S'(u))$ and $(x_2, y_2) = (S(v), S'(v))$ determine the affine line

$$Y - mX - c,$$

where

$$m = [S'(v) - S'(u)]/[S(v) - S(u)],$$
$$c = [S'(u)S(v) - S'(v)S(u)]/[S(v) - S(u)].$$

The third point of intersection of this line with the cubic $\mathscr{C}_\lambda$ is the point (x_3, y_3), where

$$\begin{aligned} x_3 &= c^2/4\lambda x_1 x_2 \\ &= [S'(u)S(v) - S'(v)S(u)]^2/4\lambda S(u)S(v)[S(v) - S(u)]^2 \\ &= 1/\lambda S(u+v). \end{aligned}$$

On the other hand, the points $(0, 0) = (S(0), S'(0))$ and $(x_3^*, y_3^*) = (S(u+v), S'(u+v))$ determine the affine line $Y - (y_3^*/x_3^*)X$ and its third point of intersection with $\mathscr{C}_\lambda$ is the point (x_4, y_4), where $x_4 = 1/\lambda x_3^* = x_3$. Evidently $y_4^2 = y_3^2$, and it may be verified that actually $y_4 = y_3$. Thus (x_3^*, y_3^*) is the third point of intersection with $\mathscr{C}_\lambda$ of the line determined by the points $(0, 0)$ and (x_3, y_3).

The origin $(0, 0)$ may not be a point of the cubic curve $\mathscr{C}_{a,b}$ but $\boldsymbol{O}$, the point at infinity, certainly is. Consequently, as illustrated in Figure 2, we now define the *sum* $P_1 + P_2$ of two K-points P_1, P_2 of $\mathscr{C}_{a,b}$ to be the K-point P_3^*, where P_3 is the third point of $\mathscr{C}_{a,b}$ on the line determined by P_1, P_2 and P_3^* is the third point of $\mathscr{C}_{a,b}$ on the line determined by $\boldsymbol{O}, P_3$. If $P_1 = P_2$, the line determined by P_1, P_2 is understood to mean the tangent to $\mathscr{C}_{a,b}$ at P_1.

It is simply a matter of elementary algebra to deduce from the formulas previously given that, if addition is defined in this way, the set of all K-points of $\mathscr{C}_{a,b}$ becomes an abelian group, with $\boldsymbol{O}$ as identity element and with $-P = (x, -y)$ as the inverse of $P = (x, y)$. Since $-P = P$ if and only if $y = 0$, the elements of order 2 in this group are the points $(x_0, 0)$, where x_0 is a root of the polynomial $X^3 + aX + b$ (if it has any roots in K).

Throughout the preceding discussion of cubic curves we restricted attention to those with a flex. It will now be shown that in a sense this is no restriction.

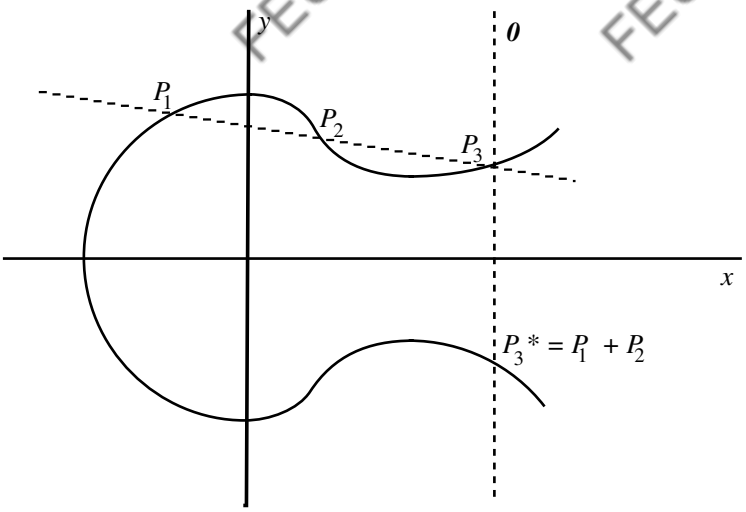

Fig. 2. Addition on $\mathscr{C}_{a,b}$.

Let $\mathscr{C}$ be a projective cubic curve over the field K, defined by the homogeneous polynomial $F_1(X, Y, Z)$, and suppose that $\mathscr{C}$ has a non-singular K-point P. Without loss of generality we assume that $P = (1, 0, 0)$ and that the tangent at P is the projective line Z. Then F_1 has no term in X^3 or in X^2Y:

$$F_1(X, Y, Z) = aY^3 + bY^2Z + cYZ^2 + dZ^3 + eX^2Z + gXY^2 + hXZ^2.$$

Here $e \neq 0$, since P is non-singular, and we may suppose $g \neq 0$, since otherwise P is a flex. If we replace $gX + aY$ by X, this assumes the form

$$F_2(X, Y, Z) = XY^2 + bY^2Z + cYZ^2 + dZ^3 + eX^2Z + gXYZ + hXZ^2,$$

with new values for the coefficients. If we now replace $X + bZ$ by X, this assumes the form

$$F_3(X, Y, Z) = XY^2 + cYZ^2 + dZ^3 + eX^2Z + gXYZ + hXZ^2,$$

again with new values for the coefficients. The projective cubic curve $\mathscr{D}$ over the field K, defined by the homogeneous polynomial

$$F_4(U, V, W) = VW^2 + cV^2W + dUV^2 + eU^3 + gUVW + hU^2V,$$

has a flex at the point $(0, 0, 1)$. Moreover,

$$F_3(U^2, VW, UV) = U^2VF_4(U, V, W),$$
$$F_4(XZ, Z^2, XY) = XZ^2F_3(X, Y, Z).$$

This shows that *any projective cubic curve over the field K with a non-singular K-point is* birationally equivalent *to one with a flex*.

Birational equivalence may be defined in the following way. A *rational transformation* of the projective plane with points $\boldsymbol{X} = (X_1, X_2, X_3)$ is a map $\boldsymbol{X} \to \boldsymbol{Y} = \boldsymbol{\varphi}(\boldsymbol{X})$, where

$$\boldsymbol{\varphi}(\boldsymbol{X}) = (\varphi_1(\boldsymbol{X}), \varphi_2(\boldsymbol{X}), \varphi_3(\boldsymbol{X}))$$

and $\varphi_1, \varphi_2, \varphi_3$ are homogeneous polynomials without common factor of the same degree m, say. (In the corresponding affine plane the coordinates are transformed by *rational* functions.) The transformation is *birational* if there exists an inverse map $\boldsymbol{Y} \to \boldsymbol{X} = \boldsymbol{\psi}(\boldsymbol{Y})$, where

$$\boldsymbol{\psi}(\boldsymbol{Y}) = (\psi_1(\boldsymbol{Y}), \psi_2(\boldsymbol{Y}), \psi_3(\boldsymbol{Y}))$$

and ψ_1, ψ_2, ψ_3 are homogeneous polynomials without common factor of the same degree n, say, such that

$$\boldsymbol{\psi}[\boldsymbol{\varphi}(\boldsymbol{X})] = \omega(\boldsymbol{X})\boldsymbol{X}, \quad \boldsymbol{\varphi}[\boldsymbol{\psi}(\boldsymbol{Y})] = \theta(\boldsymbol{Y})\boldsymbol{Y}$$

for some scalar polynomials $\omega(\boldsymbol{X}), \theta(\boldsymbol{Y})$. Two irreducible projective plane curves $\mathscr{C}$ and $\mathscr{D}$ over the field K, defined respectively by the homogeneous polynomials $F(\boldsymbol{X})$ and $G(\boldsymbol{Y})$ (not necessarily of the same degree), are *birationally equivalent* if there exists a birational transformation $\boldsymbol{Y} = \varphi(\boldsymbol{X})$ with inverse $\boldsymbol{X} = \psi(\boldsymbol{Y})$ such that $G[\varphi(\boldsymbol{X})]$ is divisible by $F(\boldsymbol{X})$ and $F[\psi(\boldsymbol{Y})]$ is divisible by $G[(\boldsymbol{Y})]$.

It is clear that birational equivalence is indeed an equivalence relation, and that irreducible projective curves which are projectively equivalent are also birationally equivalent. Birational transformations are often used to simplify the singular points of a curve. Indeed the theorem on *resolution of singularities* says that any irreducible curve is birationally equivalent to a non-singular curve, although it may be a curve in a higher-dimensional space rather than in the plane. The algebraic geometry of curves may be regarded as the study of those properties which are invariant under birational equivalence.

It was shown by Poincaré (1901) that any non-singular curve of *genus* 1 defined over the field $\mathbb{Q}$ of rational numbers and with at least one rational point is birationally equivalent over $\mathbb{Q}$ to a cubic curve. Such a curve is now said to be an *elliptic curve* (for the somewhat inadequate reason that it may be parametrized by elliptic functions over the field of complex numbers.) However, for our purposes it is sufficient to define an elliptic curve to be a non-singular cubic curve of the form $\mathscr{W}$, over a field K of arbitrary characteristic, or of the form $\mathscr{C}_{a,b}$, over a field K of characteristic $\neq 2, 3$.

4 Mordell's Theorem

We showed in the previous section that, for any field K of characteristic $\neq 2, 3$, the K-points of the elliptic curve $\mathscr{C}_{a,b}$ defined by the polynomial

$$Y^2 - X^3 - aX - b,$$

where $a, b \in K$ and $d := 4a^3 + 27b^2 \neq 0$, form an abelian group, $E(K)$ say. We now restrict our attention to the case when $K = \mathbb{Q}$ is the field of rational numbers, and

we write simply $E := E(\mathbb{Q})$. This section is devoted to the basic theorem of Mordell (1922), which says that *the abelian group E is finitely generated*.

By replacing X by X/c^2 and Y by Y/c^3 for some nonzero $c \in \mathbb{Q}$, we may (and will) assume that a and b are both integers. Let $P = (x, y)$ be any finite rational point of $\mathscr{C}_{a,b}$ and write $x = p/q$, where p and q are coprime integers. The *height* $h(P)$ of P is uniquely defined by

$$h(P) = \log \max(|p|, |q|).$$

We also set $h(\boldsymbol{O}) = 0$, where $\boldsymbol{O}$ is the unique point at infinity of $\mathscr{C}_{a,b}$.

Evidently $h(P) \geq 0$. Furthermore, $h(-P) = h(P)$, since $P = (x, y)$ implies $-P = (x, -y)$. Also, for any $r > 0$, there exist only finitely many elements $P = (x, y)$ of E with $h(P) \leq r$, since x determines y up to sign.

Proposition 8 *There exists a constant $C = C(a, b) > 0$ such that*

$$|h(2P) - 4h(P)| \leq C \quad \text{for all } P \in E.$$

Proof By the formulas given in §3, if $P = (x, y)$, then $2P = (x', y')$, where

$$x' = m^2 - 2x, \quad m = (3x^2 + a)/2y.$$

Since $y^2 = x^3 + ax + b$, it follows that

$$x' = (x^4 - 2ax^2 - 8bx + a^2)/4(x^3 + ax + b).$$

If $x = p/q$, where p and q are coprime integers, then $x' = p'/q'$, where

$$p' = p^4 - 2ap^2q^2 - 8bpq^3 + a^2q^4,$$
$$q' = 4q(p^3 + apq^2 + bq^3).$$

Evidently p' and q' are also integers, but they need not be coprime. However, since

$$p' = ep'', \quad q' = eq'',$$

where e, p'', q'' are integers and p'', q'' are coprime, we have

$$h(2P) = \log \max(|p''|, |q''|) \leq \log \max(|p'|, |q'|).$$

Since

$$\max(|p'|, |q'|) \leq \max(|p|, |q|)^4 \max\{1 + 2|a| + 8|b| + a^2, 4(1 + |a| + |b|)\},$$

it follows that

$$h(2P) \leq 4h(P) + C'$$

for some constant $C' = C'(a, b) > 0$.

The Euclidean algorithm may be used to derive the polynomial identity

$$(3X^2+4a)(X^4-2aX^2-8bX+a^2)-(3X^3-5aX-27b)(X^3+aX+b)=d,$$

where once again $d=4a^3+27b^2$. Substituting p/q for X, we obtain

$$4dq^7=4(3p^2q+4aq^3)p'-(3p^3-5apq^2-27bq^3)q'.$$

Similarly, the Euclidean algorithm may be used to derive the polynomial identity

$$f(X)(1-2aX^2-8bX^3+a^2X^4)+g(X)X(1+aX^2+bX^3)=d,$$

where

$$\begin{aligned}f(X)&=4a^3+27b^2-a^2bX+a(3a^3+22b^2)X^2+3b(a^3+8b^2)X^3,\\ g(X)&=a^2b+a(5a^3+32b^2)X+2b(13a^3+96b^2)X^2-3a^2(a^3+8b^2)X^3.\end{aligned}$$

Substituting q/p for X, we obtain

$$\begin{aligned}4dp^7=4\{&(4a^3+27b^2)p^3-a^2bp^2q+(3a^4+22ab^2)pq^2+3(a^3b+8b^3)q^3\}p'\\ &+\{a^2bp^3+(5a^4+32ab^2)p^2q+(26a^3b+192b^3)pq^2-3(a^5+8a^2b^2)q^3\}q'.\end{aligned}$$

Since $d\neq 0$, it follows from these two relations that

$$\max(|p|,|q|)^7\le C_1\max(|p|,|q|)^3\max(|p'|,|q'|)$$

and hence

$$\max(|p|,|q|)^4\le C_1\max(|p'|,|q'|).$$

But the two relations also show that the greatest common divisor e of p' and q' divides both $4dq^7$ and $4dp^7$, and hence also $4d$, since p and q are coprime. Consequently

$$\max(|p'|,|q'|)\le 4|d|\max(|p''|,|q''|).$$

Combining this with the previous inequality, we obtain

$$4h(P)\le h(2P)+C''$$

for some constant $C''=C''(a,b)>0$.

This proves the result, with $C=\max(C',C'')$. □

Proposition 9 *There exists a unique function $\hat{h}\colon E\to\mathbb{R}$ such that*

(i) *$\hat{h}-h$ is bounded,*
(ii) *$\hat{h}(2P)=4\hat{h}(P)$ for every $P\in E$.*

Furthermore, it is given by the formula $\hat{h}(P)=\lim_{n\to\infty}h(2^nP)/4^n$.

Proof Suppose $\hat{h}$ has the properties (i),(ii). Then, by (ii), $4^n\hat{h}(P) = \hat{h}(2^nP)$ and hence, by (i), $4^n\hat{h}(P) - h(2^nP)$ is bounded. Dividing by 4^n, we see that $h(2^nP)/4^n \to \hat{h}(P)$ as $n \to \infty$. This proves uniqueness.

To prove existence, choose C as in the statement of Proposition 8. Then, for any integers m, n with $n > m > 0$,

$$\begin{aligned}|4^{-n}h(2^nP) - 4^{-m}h(2^mP)| &= \left|\sum_{j=m}^{n-1}\{4^{-j-1}h(2^{j+1}P) - 4^{-j}h(2^jP)\}\right| \\ &\le \sum_{j=m}^{n-1} 4^{-j-1}|h(2^{j+1}P) - 4h(2^jP)| \\ &\le \sum_{j=m}^{n-1} 4^{-j-1}C < 4^{-m}C/3.\end{aligned}$$

Thus the sequence $\{4^{-n}h(2^nP)\}$ is a fundamental sequence and consequently convergent. If we denote its limit by $\hat{h}(P)$, then clearly $\hat{h}(2P) = 4\hat{h}(P)$. On the other hand, by taking $m = 0$ and letting $n \to \infty$ in the preceding inequality we obtain

$$|\hat{h}(P) - h(P)| \le C/3.$$

Thus $\hat{h}$ has both the required properties. □

The value $\hat{h}(P)$ is called the *canonical height* of the rational point P. The formula for $\hat{h}(P)$ shows that, for all $P \in E$,

$$\hat{h}(-P) = \hat{h}(P) \ge 0.$$

Moreover, by Proposition 9(i), for any $r > 0$ there exist only finitely many elements P of E with $\hat{h}(P) \le r$.

It will now be shown that the canonical height satisfies the *parallelogram law*:

Proposition 10 *For all $P_1, P_2 \in E$,*

$$\hat{h}(P_1 + P_2) + \hat{h}(P_1 - P_2) = 2\hat{h}(P_1) + 2\hat{h}(P_2).$$

Proof It is sufficient to show that there exists a constant $C' > 0$ such that, for all $P_1, P_2 \in E$,

$$h(P_1 + P_2) + h(P_1 - P_2) \le 2h(P_1) + 2h(P_2) + C'. \qquad (*)$$

For it then follows from the formula in Proposition 9 that, for all $P_1, P_2 \in E$,

$$\hat{h}(P_1 + P_2) + \hat{h}(P_1 - P_2) \le 2\hat{h}(P_1) + 2\hat{h}(P_2).$$

But, replacing P_1 by $P_1 + P_2$ and P_2 by $P_1 - P_2$, we also have

$$\hat{h}(2P_1) + \hat{h}(2P_2) \le 2\hat{h}(P_1 + P_2) + 2\hat{h}(P_1 - P_2)$$

and hence, by Proposition 9(ii),

$$2\hat{h}(P_1) + 2\hat{h}(P_2) \le \hat{h}(P_1 + P_2) + \hat{h}(P_1 - P_2).$$

To prove (∗) we may evidently assume that $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ are both finite. Moreover, by Proposition 8, we may assume that $P_1 \neq P_2$. Then, by the formulas of §3,

$$P_1 + P_2 = (x_3, y_3), \quad P_1 - P_2 = (x_4, y_4),$$

where

$$x_3 = (y_2 - y_1)^2/(x_2 - x_1)^2 - (x_1 + x_2),$$
$$x_4 = (y_2 + y_1)^2/(x_2 - x_1)^2 - (x_1 + x_2).$$

Hence

$$x_3 + x_4 = 2[y_2^2 + y_1^2 - (x_2 - x_1)(x_2^2 - x_1^2)]/(x_2 - x_1)^2$$

and

$$x_3 x_4 = (y_2^2 - y_1^2)^2/(x_2 - x_1)^4 - 2(x_1 + x_2)(y_1^2 + y_2^2)/(x_2 - x_1)^2 + (x_1 + x_2)^2.$$

Since $y_j^2 = x_j^3 + ax_j + b$ $(j = 1, 2)$, these relations simplify to

$$x_3 + x_4 = 2[x_1 x_2 (x_1 + x_2) + a(x_1 + x_2) + 2b]/(x_2 - x_1)^2$$

and

$$x_3 x_4 = N/(x_2 - x_1)^2,$$

where

$$\begin{aligned} N &= (x_2^2 + x_1 x_2 + x_1^2 + a)^2 - 2(x_1 + x_2)^2(x_2^2 - x_1 x_2 + x_1^2 + a) \\ &\quad - 4b(x_1 + x_2) + (x_2^2 - x_1^2)^2 \\ &= (x_1 x_2 - a)^2 - 4b(x_1 + x_2). \end{aligned}$$

Put $x_j = p_j/q_j$, where $(p_j, q_j) = 1$ $(1 \leq j \leq 4)$. Then x_3, x_4 are the roots of the quadratic polynomial

$$AX^2 + BX + C$$

with integer coefficients

$$\begin{aligned} A &= (p_2 q_1 - p_1 q_2)^2, \\ B &= (p_1 p_2 + a q_1 q_2)(p_1 q_2 + p_2 q_1) + 2b q_1^2 q_2^2, \\ C &= (p_1 p_2 - a q_1 q_2)^2 - 4b q_1 q_2 (p_1 q_2 + p_2 q_1). \end{aligned}$$

Consequently

$$\begin{aligned} A p_3 p_4 &= C q_3 q_4, \\ A(p_3 q_4 + p_4 q_3) &= B q_3 q_4. \end{aligned}$$

By Proposition II.16, q_3 and q_4 each divide A, and so their product divides A^2. Hence, for some integer $D \neq 0$,

$$A^2 = Dq_3q_4, \quad AC = Dp_3p_4, \quad AB = D(p_3q_4 + p_4q_3).$$

But it is easily seen that q_3q_4, p_3p_4 and $p_3q_4 + p_4q_3$ have no common prime divisor. It follows that A divides D.

Hence, if we put

$$\rho_j = \max(|p_j|, |q_j|) \ (1 \leq j \leq 4),$$

then

$$\begin{aligned} |q_3q_4| \leq |A| &\leq 4\rho_1^2\rho_2^2, \\ |p_3p_4| \leq |C| &\leq [(1 + |a|)^2 + 8|b|]\rho_1^2\rho_2^2, \\ |p_3q_4 + p_4q_3| \leq |B| &\leq 2(1 + |a| + |b|)\rho_1^2\rho_2^2. \end{aligned}$$

But

$$\max(|p_3|, |q_3|) \max(|p_4|, |q_4|) \leq \max(|p_3p_4|, |q_3q_4| + |p_3q_4 + p_4q_3|),$$

since if $|q_3| \leq |p_3|$ and $|p_4| \leq |q_4|$, for example, then

$$|p_3q_4| \leq |p_4q_3| + |p_3q_4 + p_4q_3| \leq |q_3q_4| + |p_3q_4 + p_4q_3|.$$

It follows that there exists a constant $C'' > 0$ such that

$$\rho_3\rho_4 \leq C''\rho_1^2\rho_2^2,$$

which is equivalent to $(*)$ with $C' = \log C''$. □

Corollary 11 *For any $P \in E$ and any integer n,*

$$\hat{h}(nP) = n^2\hat{h}(P).$$

Proof Since $\hat{h}(-P) = \hat{h}(P)$, we may assume $n > 0$. We may actually assume $n > 2$, since the result is trivial for $n = 1$ and it holds for $n = 2$ by Proposition 9. By Proposition 10 we have

$$\hat{h}(nP) + \hat{h}((n-2)P) = 2\hat{h}((n-1)P) + 2\hat{h}(P),$$

from which the general case follows by induction. □

It follows from Corollary 11 that if an element P of the group E has finite order, then $\hat{h}(P) = 0$. The converse is also true. In fact, by Proposition 10, the set of all $P \in E$ such that $\hat{h}(P) = 0$ is a subgroup of E, and this subgroup is finite since there are only finitely many points P such that $\hat{h}(P) < 1$.

We now deduce from Proposition 10 that a non-negative quadratic form can be constructed from the canonical height. If we put

$$(P, Q) = \hat{h}(P + Q) - \hat{h}(P) - \hat{h}(Q),$$

then evidently

$$(P, Q) = (Q, P), (P, P) = 2\hat{h}(P) \geq 0.$$

It remains to show that

$$(P, Q + R) = (P, Q) + (P, R),$$

and we do this by proving that

$$\hat{h}(P + Q + R) = \hat{h}(P + Q) + \hat{h}(P + R) + \hat{h}(Q + R) - \hat{h}(P) - \hat{h}(Q) - \hat{h}(R).$$

But, by the parallelogram law,

$$\begin{aligned}\hat{h}(P + Q + R + P) + \hat{h}(Q + R) &= \hat{h}(P + Q + R + P) + \hat{h}(P + Q + R - P)\\ &= 2\hat{h}(P + Q + R) + 2\hat{h}(P)\end{aligned}$$

and

$$\begin{aligned}\hat{h}(P + Q + R + P) + \hat{h}(Q - R) &= \hat{h}(P + Q + R + P) + \hat{h}(P + Q - P - R)\\ &= 2\hat{h}(P + Q) + 2\hat{h}(P + R).\end{aligned}$$

Subtracting the second relation from the first, we obtain

$$\hat{h}(Q + R) - \hat{h}(Q - R) = 2\hat{h}(P + Q + R) + 2\hat{h}(P) - 2\hat{h}(P + Q) - 2\hat{h}(P + R).$$

Since, by the parallelogram law again,

$$\hat{h}(Q + R) + \hat{h}(Q - R) = 2\hat{h}(Q) + 2\hat{h}(R),$$

this is equivalent to what we wished to prove.

Proposition 12 *The abelian group E is finitely generated if, for some integer $m > 1$, the factor group E/mE is finite.*

Proof Let S be a set of representatives of the cosets of the subgroup mE. Since S is finite, by hypothesis, we can choose $C > 0$ so that $\hat{h}(Q) \leq C$ for all $Q \in S$. The set

$$S' = \{Q' \in E : \hat{h}(Q') \leq C\}$$

contains S and is also finite. We will show that it generates E.

Let E' be the subgroup of E generated by the elements of S'. If $E' \neq E$, choose $P \in E \backslash E'$ so that $\hat{h}(P)$ is minimal. Then

$$P = mP_1 + Q_1 \quad \text{for some } P_1 \in E \text{ and } Q_1 \in S.$$

Since

$$\hat{h}(P + Q_1) + \hat{h}(P - Q_1) = 2\hat{h}(P) + 2\hat{h}(Q_1),$$

it follows that

$$\hat{h}(mP_1) = \hat{h}(P - Q_1) \leq 2\hat{h}(P) + 2C$$

and hence

$$\hat{h}(P_1) \leq 2[\hat{h}(P) + C]/m^2 \leq [\hat{h}(P) + C]/2.$$

But $P_1 \notin E'$, since $P \notin E'$, and hence $\hat{h}(P_1) \geq \hat{h}(P)$. It follows that $\hat{h}(P) \leq C$, which is a contradiction. Hence $E' = E$. □

Proposition 12 shows that to complete the proof of Mordell's theorem it is enough to show that the factor group $E/2E$ is finite. *We will prove this only for the case when E contains an element of order* 2. A similar proof may be given for the general case, but it requires some knowledge of algebraic number theory.

The assumption that E contains an element of order 2 means that there is a rational point $(x_0, 0)$, where x_0 is a root of the polynomial $X^3 + aX + b$. Since a and b are taken to be integers, and the polynomial has highest coefficient 1, x_0 must also be an integer. By changing variable from X to $x_0 + X$, we replace the cubic $\mathscr{C}_{a,b}$ by a cubic $C_{A,B}$ defined by a polynomial

$$Y^2 - (X^3 + AX^2 + BX),$$

where $A, B \in \mathbb{Z}$. The non-singularity condition $d := 4a^3 + 27b^2 \neq 0$ becomes

$$D := B^2(4B - A^2) \neq 0,$$

but this is the only restriction on A, B. The chord joining two rational points of $C_{A,B}$ is given by the same formulas as for $\mathscr{C}_{a,b}$ in §3, but the tangent to $C_{A,B}$ at the finite point $P_1 = (x_1, y_1)$ is now the affine line

$$Y - mX - c,$$

where

$$m = (3x_1^2 + 2Ax_1 + B)/2y_1, \quad c = -x_1(x_1^2 - B)/2y_1.$$

The geometrical interpretation of the group law remains the same as before. We will now denote by E the group of all rational points of $C_{A,B}$. Our change of variable has made the point $\boldsymbol{N} = (0, 0)$ an element of E of order 2.

Let $P = (x, y)$ be a rational point of $C_{A,B}$ with $x \neq 0$. We are going to show that, in a sense which will become clear, there are only finitely many rational square classes to which x can belong.

Write $x = m/n$, $y = p/q$, where m, n, p, q are integers with $n, q > 0$ and $(m, n) = (p, q) = 1$. Then

$$p^2n^3 = (m^3 + Am^2n + Bmn^2)q^2,$$

which implies both $q^2 | n^3$ and $n^3 | q^2$. Thus $n^3 = q^2$. From $n^2 | q^2$ we obtain $n | q$. Hence $q = en$ for some integer e, and it follows that $n = e^2$, $q = e^3$. Thus

$$x = m/e^2, \quad y = p/e^3, \quad \text{where } e > 0 \quad \text{and} \quad (m, e) = (p, e) = 1.$$

Moreover,

$$p^2 = m(m^2 + Ame^2 + Be^4).$$

This shows that each prime which divides m, but not $m^2 + Ame^2 + Be^4$, must occur to an even power in m. On the other hand, each prime which divides both m and $m^2 + Ame^2 + Be^4$ must also divide B, since $(m, e) = 1$. Consequently we can write

$$x = \pm p_1^{\varepsilon_1} \cdots p_k^{\varepsilon_k} (u/e)^2,$$

where $u \in \mathbb{N}$, $p_1, \ldots, p_k$ are the distinct primes dividing B and $\varepsilon_j \in \{0, 1\}$ $(1 \leq j \leq k)$. Hence there are at most 2^{k+1} rational square classes to which x can belong.

Suppose now that $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ are distinct rational points of $C_{A,B}$ for which $x_1 x_2$ is a nonzero rational square, and let $P_3 = (x_3, y_3)$ be the third point of intersection with $C_{A,B}$ of the line through P_1 and P_2. Then x_1, x_2, x_3 are the three roots of a cubic equation

$$(mX + c)^2 = X^3 + AX^2 + BX.$$

From the constant term we see that $x_1 x_2 x_3 = c^2$. It follows that x_3 is a nonzero rational square if $c \neq 0$. If $c = 0$, then $P_3 = N$ and $x_1 x_2 = B$.

Suppose next that $P = (x, y)$ is any rational point of $C_{A,B}$ with $x \neq 0$, and let $2P = (\bar{x}, -\bar{y})$. Then $\bar{P} = (\bar{x}, \bar{y})$ is the other point of intersection with $C_{A,B}$ of the tangent to $C_{A,B}$ at P. By the same argument as before, $x^2 \bar{x} = c^2$. Hence $\bar{x}$ is a nonzero rational square if $c \neq 0$. If $c = 0$, then $2P = N$ and $x^2 = B$.

To deduce that $E/2E$ is finite from these observations we will use an arithmetic analogue of Landen's transformation. We saw in Chapter XII that, over the field $\mathbb{C}$ of complex numbers, the cubic curve $\mathscr{C}_\lambda$ defined by the polynomial $Y^2 - g_\lambda(X)$, where $g_\lambda(X) = 4X(1 - X)(1 - \lambda X)$, admits the parametrization

$$X = S(u, \lambda), \quad Y = S'(u, \lambda).$$

It follows from Proposition XII.11 that the cubic curve $\mathscr{C}_{\lambda'}$, where λ' is given by $\lambda' = \lambda^2/[1 + (1 - \lambda)^{1/2}]^4$, admits the parametrization

$$X' = [1 + (1 - \lambda)^{1/2}]X(1 - X)/(1 - \lambda X),$$
$$Y' = [1 + (1 - \lambda)^{1/2}]Y(1 - 2X + \lambda X^2)/(1 - \lambda X)^2,$$

where again $X = S(u, \lambda)$, $Y = S'(u, \lambda)$ and where $(1 - 2X + \lambda X^2)/(1 - \lambda X)^2$ is the derivative with respect to X of $X(1 - X)/(1 - \lambda X)$. Since also $X' = S(u', \lambda')$, where $u' = [1 + (1 - \lambda)^{1/2}]u$, the map $(X, Y) \to (X', Y')$ defines a homomorphism of the group of complex points of $\mathscr{C}_\lambda$ into the group of complex points of $\mathscr{C}_{\lambda'}$.

We will simply state analogous results for the cubic curve $C_{A,B}$ over the field $\mathbb{Q}$ of rational numbers, since their verification is elementary. If (x, y) is a rational point of $C_{A,B}$ with $x \neq 0$ and if

$$x' = (x^2 + Ax + B)/x, \quad y' = y(x^2 - B)/x^2,$$

then (x', y') is a rational point of $C_{A',B'}$, where

$$A' = -2A, \quad B' = A^2 - 4B.$$

Moreover, if we define a map φ of the group E of all rational points of $C_{A,B}$ into the group E' of all rational points of $C_{A',B'}$ by putting

$$\varphi(x, y) = (x', y') \quad \text{if } x \neq 0, \quad \varphi(\boldsymbol{N}) = \varphi(\boldsymbol{O}) = \boldsymbol{O},$$

then φ is a homomorphism, i.e.

$$\varphi(P + Q) = \varphi(P) + \varphi(Q), \quad \varphi(-P) = -\varphi(P).$$

The range $\varphi(E)$ may not be the whole of E'. In fact, since

$$x' = (x^3 + Ax^2 + Bx)/x^2 = (y/x)^2,$$

the first coordinate of any finite point of $\varphi(E)$ must be a rational square. Furthermore, if $\boldsymbol{N} = (0, 0)$ is a point of $\varphi(E)$, the integer $B' = A^2 - 4B$ must be a square. We will show that these conditions completely characterize $\varphi(E)$.

Evidently if $A^2 - 4B$ is a square, then the quadratic polynomial $X^2 + AX + B$ has a rational root $x_0 \neq 0$ and $\varphi(x_0, 0) = \boldsymbol{N}$. Suppose now that (x', y') is a rational point of $C_{A',B'}$ and that $x' = t^2$ is a nonzero rational square. We will show that if

$$x_1 = (t^2 - A + y'/t)/2, \quad y_1 = tx_1,$$
$$x_2 = (t^2 - A - y'/t)/2, \quad y_2 = -tx_2,$$

then $(x_j, y_j) \in E$ and $\varphi(x_j, y_j) = (x', y')$ $(j = 1, 2)$. It is easily seen that $(x_j, y_j) \in E$ if and only if

$$t^2 = x_j + A + B/x_j.$$

But

$$\begin{aligned} x_1x_2 &= [(t^2 - A)^2 - y'^2/t^2]/4 \\ &= [(x' - A)^2 - y'^2/x']/4 \\ &= (x'^3 - 2Ax'^2 + A^2x' - y'^2)/4x'. \end{aligned}$$

Since

$$y'^2 = x'^3 - 2Ax'^2 + (A^2 - 4B)x',$$

it follows that $x_1x_2 = B$. Hence (x_1, y_1) and (x_2, y_2) are both in E if $t^2 = x_1 + A + x_2$, and this condition is certainly satisfied by the definitions of x_1 and x_2.

In addition to

$$x_j + A + B/x_j = t^2 = x' \ (j = 1, 2),$$

we have

$$y_1(x_1^2 - B)/x_1^2 = t(x_1^2 - x_1x_2)/x_1 = t(x_1 - x_2) = y',$$

and similarly $y_2(x_2^2 - B)/x_2^2 = y'$. It follows that

$$\varphi(x_1, y_1) = \varphi(x_2, y_2) = (x', y').$$

Since φ is a homomorphism, the range $\varphi(E)$ is a subgroup of E'. We are going to show that this subgroup is of finite index in E'. By what we have already proved for E, there exists a finite (or empty) set $P_1' = (x_1', y_1'), \ldots, P_s' = (x_s', y_s')$ of points of E' such that x_i' is not a rational square ($1 \le i \le s$) and such that, if $P' = (x', y')$ is any other point of E' for which x' is not a rational square, then $x'x_j'$ is a nonzero rational square for a unique $j \in \{1, \ldots, s\}$. Let $P'' = (x'', y'')$ be the third point of intersection with $C_{A',B'}$ of the line through P' and P_j', so that

$$P' + P_j' + P'' = \boldsymbol{O}.$$

By what we have already proved, either x'' is a nonzero rational square or $P'' = \boldsymbol{N}$ and $x'x_j' = B'$ is a square. In either case, $P'' \in \varphi(E)$. Furthermore, if $2P_j' = (\bar{x}, -\bar{y})$, then either $\bar{x}$ is a nonzero rational square or $2P_j' = \boldsymbol{N}$ and $x_j'^2 = B'$. In either case again, $2P_j' \in \varphi(E)$. Since

$$P' = P_j' - (2P_j' + P''),$$

it follows that P' and P_j' are in the same coset of $\varphi(E)$. Consequently $P_1', \ldots, P_s'$, together with $\boldsymbol{O}$, and also $\boldsymbol{N}$ if B' is not a square, form a complete set of representatives of the cosets of $\varphi(E)$ in E'.

The preceding discussion can be repeated with $C_{A',B'}$ in the place of $C_{A,B}$. It yields a homomorphism φ' of the group E' of all rational points of $C_{A',B'}$ into the group E'' of all rational points of $C_{A'',B''}$, where

$$A'' = -2A' = 4A, \quad B'' = A'^2 - 4B' = 16B.$$

But the simple transformation $(X, Y) \to (X/4, Y/8)$ replaces $C_{A'',B''}$ by $C_{A,B}$ and defines an isomorphism χ of E'' with E. Hence the composite map $\psi = \chi \circ \varphi'$ is a homomorphism of E' into E, and $\psi \circ \varphi$ is a homomorphism of E into itself.

We now show that the homomorphism $P \to \psi \circ \varphi(P)$ is just the doubling map $P \to 2P$. Since this is obvious if $P = \boldsymbol{O}$ or $\boldsymbol{N}$, we need only verify it for $P = (x, y)$ with $x \ne 0$.

For $P'' = \varphi' \circ \varphi(P)$ we have

$$x'' = (y'/x')^2 = [y(1 - B/x^2) \cdot x^2/y^2]^2 = (x^2 - B)^2/y^2$$

and

$$\begin{aligned} y'' &= y'(1 - B'/x'^2) = y(1 - B/x^2)[1 - (A^2 - 4B)x^4/y^4] \\ &= (x^2 - B)[y^4 - (A^2 - 4B)x^4]/x^2y^3 \\ &= (x^2 - B)[(x^2 + Ax + B)^2 - (A^2 - 4B)x^2]/y^3. \end{aligned}$$

Hence for $\psi \circ \varphi(P) = P^* = (x^*, y^*)$ we have

$$x^* = (x^2 - B)^2/4y^2,$$
$$y^* = (x^2 - B)[(x^2 + Ax + B)^2 - (A^2 - 4B)x^2]/8y^3.$$

On the other hand, if the tangent to $C_{A,B}$ at P intersects $C_{A,B}$ again at $(\bar{x}, \bar{y})$, then $2P = (\bar{x}, -\bar{y})$. The cubic equation

$$(mx + c)^2 = X^3 + AX^2 + BX$$

has x as a double root and $\bar{x}$ as its third root. Hence $\bar{x} = (c/x)^2$. Using the formula for c given previously, we obtain

$$\bar{x} = (x^2 - B)^2/4y^2 = x^*.$$

Furthermore, using the formula for m given previously,

$$\bar{y} = m\bar{x} + c = [(3x^2 + 2Ax + B)\bar{x} - x(x^2 - B)]/2y$$
$$= (x^2 - B)[(3x^2 + 2Ax + B)(x^2 - B) - 4xy^2)]/8y^3.$$

Substituting $x^3 + Ax^2 + Bx$ for y^2, we obtain $\bar{y} = -y^*$. Thus $\psi \circ \varphi(P) = 2P$, as claimed.

Since $\varphi(E)$ has finite index in E', and likewise $\psi(E')$ has finite index in E, it follows that $2E = \psi \circ \varphi(E)$ has finite index in E. (The proof shows that the index is at most $2^{\alpha+\beta+2}$, where α is the number of distinct prime divisors of B and β is the number of distinct prime divisors of $A^2 - 4B$.)

By the remarks after the proof of Proposition 12, Mordell's theorem has now been completely proved in the case where E contains an element of order 2.

5 Further Results and Conjectures

Let $\mathscr{C}_{a,b}$ be the elliptic curve defined by the polynomial

$$Y^2 - (X^3 + aX + b),$$

where $a, b \in \mathbb{Z}$ and $d := 4a^3 + 27b^2 \neq 0$. By Mordell's theorem, the abelian group $E = E_{a,b}(\mathbb{Q})$ of all rational points of $\mathscr{C}_{a,b}$ is finitely generated. It follows from the structure theorem for finitely generated abelian groups (Chapter III, §4) that E is the direct sum of a finite abelian group E^t and a 'free' abelian group E^f, which is the direct sum of $r \geq 0$ infinite cyclic subgroups. The non-negative integer r is called the *rank* of the elliptic curve and E^t its *torsion group*.

The torsion group can, in principle, be determined by a finite amount of computation. A theorem of Nagell (1935) and Lutz (1937) says that if $P = (x, y)$ is a point of E of finite order, then x and y are integers and either $y = 0$ or y^2 divides d. Thus there are only finitely many possibilities to check.

A deep theorem of Mazur (1977) says that the torsion group must be one of the following:

(i) a cyclic group of order n ($1 \leq n \leq 10$ or $n = 12$),
(ii) the direct sum of a cyclic group of order 2 and a cyclic group of order $2n$ ($1 \leq n \leq 4$).

It was already known that each of these possibilities occurs. It is easy to check if the torsion group is of type (i) or type (ii), since in the latter case there are three elements of order 2, whereas in the former case there is at most one. Mazur's result shows that an element has infinite order, if it does not have order ≤ 12.

It is conjectured that there exist elliptic curves over $\mathbb{Q}$ with arbitrarily large rank. (Examples are known of elliptic curves with rank ≥ 22.) At present no infallible algorithm is known for determining the rank of an elliptic curve, let alone a basis for the torsion-free group E^f. However, Manin (1971) devised a conditional algorithm, based on the strong conjecture of Birch and Swinnerton-Dyer which will be mentioned later. This conjecture is still unproved, but is supported by much numerical evidence.

An important way of obtaining arithmetic information about an elliptic curve is by reduction modulo a prime p. We regard the coefficients not as integers, but as integers mod p, and we look not for $\mathbb{Q}$-points, but for $\mathbb{F}_p$-points. Since the normal form $\mathscr{C}_{a,b}$ was obtained by assuming that the field had characteristic $\neq 2, 3$, we now adopt a more general normal form.

Let $\mathscr{W} = \mathscr{W}(a_1, \ldots, a_6)$ be the projective completion of the affine cubic curve defined by the polynomial

$$Y^2 + a_1XY + a_3Y - (X^3 + a_2X^2 + a_4X + a_6),$$

where $a_j \in \mathbb{Q}$ ($j = 1, 2, 3, 4, 6$). It may be shown that $\mathscr{W}$ is non-singular if and only if the *discriminant* $\Delta \neq 0$, where

$$\Delta = -b_2^2b_8 - 8b_4^3 - 27b_6^2 + 9b_2b_4b_6$$

and

$$\begin{aligned} b_2 &= a_1^2 + 4a_2, \\ b_4 &= a_1a_3 + 2a_4, \\ b_6 &= a_3^2 + 4a_6, \\ b_8 &= a_1^2a_6 - a_1a_3a_4 + 4a_2a_6 + a_2a_3^2 - a_4^2. \end{aligned}$$

(We retain the name 'discriminant', although $\Delta = -16d$ for $\mathscr{W} = \mathscr{C}_{a,b}$.) The definition of addition on $\mathscr{W}$ has the same geometrical interpretation as on $\mathscr{C}_{a,b}$, although the corresponding algebraic formulas are different. They are written out in §7.

For any $u, r, s, t \in \mathbb{Q}$ with $u \neq 0$, the invertible linear change of variables

$$X = u^2X' + r, \quad Y = u^3Y' + su^2X' + t$$

replaces $\mathscr{W}$ by a curve $\mathscr{W}'$ of the same form with discriminant $\Delta' = u^{-12}\Delta$. By means of such a transformation we may assume that the coefficients a_j are integers and that Δ,

which is now an integer, has minimal absolute value. (It has been proved by Tate that we then have $|\Delta| > 1$.) The discussion which follows presupposes that $\mathscr{W}$ is chosen in this way so that, in particular, discriminant means 'minimal discriminant'. We say that such a $\mathscr{W}$ is a *minimal model* for the elliptic curve.

For any prime p, let $\mathscr{W}_p$ be the cubic curve defined over the finite field $\mathbb{F}_p$ by the polynomial

$$Y^2 + \tilde{a}_1 XY + \tilde{a}_3 Y - (X^3 + \tilde{a}_2 X^2 + \tilde{a}_4 X + \tilde{a}_6),$$

where $\tilde{a}_j \in a_j + p\mathbb{Z}$. If $p \nmid \Delta$ the cubic curve $\mathscr{W}_p$ is non-singular, but if $p|\Delta$ then $\mathscr{W}_p$ has a unique singular point. The singular point (x_0, y_0) of $\mathscr{W}_p$ is a *cusp* if, on replacing X and Y by $x_0 + X$ and $y_0 + Y$, we obtain a polynomial of the form

$$c(aX + bY)^2 + \cdots,$$

where $a, b, c \in \mathbb{F}_p$ and the unwritten terms are of degree > 2. Otherwise, the singular point is a *node*.

For any prime p, let N_p denote the number of $\mathbb{F}_p$-points of $\mathscr{W}_p$, including the point at infinity $\boldsymbol{O}$, and put

$$c_p = p + 1 - N_p.$$

It was conjectured by Artin (1924), and proved by Hasse (1934), that

$$|c_p| \leq 2p^{1/2} \quad \text{if } p \nmid \Delta.$$

Since $2p^{1/2}$ is not an integer, this inequality says that the quadratic polynomial

$$1 - c_p T + pT^2$$

has conjugate complex roots $\gamma_p, \bar{\gamma}_p$ of absolute value $p^{-1/2}$ or, if we put $T = p^{-s}$, that the zeros of

$$1 - c_p p^{-s} + p^{1-2s}$$

lie on the line $\mathscr{R}s = 1/2$. Thus it is an analogue of the Riemann hypothesis on the zeros of $\zeta(s)$, but differs from it by having been proved. (As mentioned in §5 of Chapter IX, Hasse's result was considerably generalized by Weil (1948) and Deligne (1974).)

The *L-function* of the original elliptic curve $\mathscr{W}$ is defined by

$$L(s) = L(s, \mathscr{W}) := \prod_{p|\Delta} (1 - c_p p^{-s})^{-1} \prod_{p \nmid \Delta} (1 - c_p p^{-s} + p^{1-2s})^{-1}.$$

The first product on the right side has only finitely many factors. The infinite second product is convergent for $\mathscr{R}s > 3/2$, since

$$1 - c_p p^{-s} + p^{1-2s} = (p^{1/2-s} - p^{1/2}\gamma_p)(p^{1/2-s} - p^{1/2}\bar{\gamma}_p)$$

and $|\gamma_p| = |\bar{\gamma}_p| = p^{-1/2}$. Multiplying out the products, we obtain for $\mathscr{R}s > 3/2$ an absolutely convergent Dirichlet series

$$L(s) = \sum_{n \geq 1} c_n n^{-s}$$

with integer coefficients c_n. (If $n = p$ is prime, then c_n is the previously defined c_p.)

The *conductor* $N = N(\mathscr{W})$ of the elliptic curve $\mathscr{W}$ is defined by the singular reductions $\mathscr{W}_p$ of $\mathscr{W}$:

$$N = \prod_{p|\Delta} p^{f_p},$$

where $f_p = 1$ if $\mathscr{W}_p$ has a node, whereas $f_p = 2$ if $p > 3$ and $\mathscr{W}_p$ has a cusp. We will not define f_p if $p \in \{2, 3\}$ and $\mathscr{W}_p$ has a cusp, but we mention that f_p is then an integer ≥ 2 which can be calculated by an algorithm due to Tate (1975). (It may be shown that $f_2 \leq 8$ and $f_3 \leq 5$.)

The elliptic curve $\mathscr{W}$ is said to be *semi-stable* if $\mathscr{W}_p$ has a node for every $p|\Delta$. Thus, for a semi-stable elliptic curve, the conductor N is precisely the product of the distinct primes dividing the discriminant Δ. (The semi-stable case is the only one in which the conductor is square-free.)

Three important conjectures about elliptic curves, involving their L-functions and conductors, will now be described.

It was conjectured by Hasse (1954) that the function

$$\zeta(s, \mathscr{W}) := \zeta(s)\zeta(s-1)/L(s, \mathscr{W})$$

may be analytically continued to a function which is meromorphic in the whole complex plane and that $\zeta(2 - s, \mathscr{W})$ is connected with $\zeta(s, \mathscr{W})$ by a functional equation similar to that satisfied by the Riemann zeta-function $\zeta(s)$. In terms of L-functions, Hasse's conjecture was given the following precise form by Weil (1967):

HW-Conjecture: *If the elliptic curve $\mathscr{W}$ has L-function $L(s)$ and conductor N, then $L(s)$ may be analytically continued, so that the function*

$$\Lambda(s) = (2\pi)^{-s}\Gamma(s)L(s),$$

where $\Gamma(s)$ denotes Euler's gamma-function, is holomorphic throughout the whole complex plane and satisfies the functional equation

$$\Lambda(s) = \pm N^{1-s}\Lambda(2-s).$$

(In fact it is the functional equation which determines the precise definition of the conductor.)

The second conjecture, due to Birch and Swinnerton-Dyer (1965), connects the L-function with the group of rational points:

BSD-Conjecture: *The L-function $L(s)$ of the elliptic curve $\mathscr{W}$ has a zero at $s = 1$ of order exactly equal to the rank $r \geq 0$ of the group $E = E(\mathscr{W}, \mathbb{Q})$ of all rational points of $\mathscr{W}$.*

This is sometimes called the 'weak' conjecture of Birch and Swinnerton-Dyer, since they also gave a 'strong' version, in which the nonzero constant C such that

$$L(s) \sim C(s-1)^r \quad \text{for } s \to 1$$

is expressed by other arithmetic invariants of $\mathscr{W}$. The strong conjecture may be regarded as an analogue for elliptic curves of a known formula for the Dedekind zeta-function of an algebraic number field. An interesting reformulation of the strong form has been given by Bloch (1980).

The statement of the third conjecture requires some preparation. For any positive integer N, let $\Gamma_0(N)$ denote the multiplicative group of all matrices

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

where a, b, c, d are integers such that $ad - bc = 1$ and $c \equiv 0 \bmod N$. A function $f(\tau)$ which is holomorphic for $\tau \in \mathscr{H}$ (the upper half-plane) is said to be a *modular form of weight* 2 *for* $\Gamma_0(N)$ if, for every such A,

$$f((a\tau + b)/(c\tau + d)) = (c\tau + d)^2 f(\tau).$$

An elliptic curve $\mathscr{W}$, with L-function

$$L(s) = \sum_{n \geq 1} c_n n^{-s}$$

and conductor N, is said to be *modular* if the function

$$f(\tau) = \sum_{n \geq 1} c_n e^{2\pi i n \tau},$$

which is certainly holomorphic in $\mathscr{H}$, is a modular form of weight 2 for $\Gamma_0(N)$. This actually implies that f is a 'cusp form' and satisfies a functional equation

$$f(-1/N\tau) = \mp N\tau^2 f(\tau).$$

It follows that the *Mellin transform*

$$\Lambda(s) = \int_0^\infty f(iy) y^{s-1} dy$$

may be analytically continued for all $s \in \mathbb{C}$ and satisfies the functional equation

$$\Lambda(s) = \pm N^{1-s} \Lambda(2-s).$$

(Note the reversal of sign.) But

$$\Lambda(s) = (2\pi)^{-s} \Gamma(s) L(s),$$

since, by (9) of Chapter IX,

$$\int_0^\infty e^{-2\pi n y} y^{s-1} dy = (2\pi n)^{-s} \Gamma(s).$$

Hence any modular elliptic curve satisfies the *HW*-conjecture.

It was shown by Weil (1967) that, conversely, an elliptic curve is modular if not only its L-function $L(s) = \sum_{n \geq 1} c_n n^{-s}$ has the properties required in the

HW-conjecture but also, for sufficiently many Dirichlet characters χ, the 'twisted' L-functions

$$L(s, \chi) = \sum_{n \geq 1} \chi(n) c_n n^{-s}$$

have analogous properties.

The definition of modular elliptic curve can be given a more intuitive form: the elliptic curve $\mathscr{C}_{a,b}$ is modular if there exist non-constant functions $X = f(\tau), Y = g(\tau)$ which are holomorphic in the upper half-plane, which are invariant under $\Gamma_0(N)$, i.e.

$$f((a\tau + b)/(c\tau + d)) = f(\tau), \quad g((a\tau + b)/(c\tau + d)) = g(\tau)$$

for every

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(N),$$

and which parametrize $\mathscr{C}_{a,b}$:

$$g^2(\tau) = f^3(\tau) + af(\tau) + b.$$

The significance of modular elliptic curves is that one can apply to them the extensive analytic theory of modular forms. For example, through the work of Kolyvagin (1990), together with results of Gross and Zagier (1986) and others, it is known that (as the *BSD*-conjecture predicts) a modular elliptic curve has rank 0 if its L-function does not vanish at $s = 1$, and has rank 1 if its L-function has a simple zero at $s = 1$.

The third conjecture, stated rather roughly by Taniyama (1955) and more precisely by Weil (1967), is simply this:

TW-Conjecture: *Every elliptic curve over the field* $\mathbb{Q}$ *of rational numbers is modular.*

The name of Shimura is often also attached to this conjecture, since he certainly contributed to its ultimate formulation. Shimura (1971) further showed that any elliptic curve which admits complex multiplication is modular. A big step forward was made by Wiles (1995) who, with assistance from Taylor, showed that any semi-stable elliptic curve is modular. A complete proof of the *TW*-conjecture, due to Diamond and others, has recently been announced by Darmon (1999). Thus all the results which had previously been established for modular elliptic curves actually hold for all elliptic curves over $\mathbb{Q}$.

It should be mentioned that there is also a 'Riemann hypothesis' for elliptic curves over $\mathbb{Q}$, namely that all zeros of the L-function in the critical strip $1/2 < \mathscr{R}s < 3/2$ lie on the line $\mathscr{R}s = 1$.

Mordell's theorem was extended from elliptic curves over $\mathbb{Q}$ to abelian varieties over any algebraic number field by Weil (1928). Many other results in the arithmetic of elliptic curves have been similarly extended. The topic is too vast to be considered here, but it should be said that our exposition for the prototype case is not always in the most appropriate form for such generalizations.

In the same paper in which he proved his theorem, Mordell (1922) conjectured that if a non-singular irreducible projective curve, defined by a homogeneous polynomial $F(x, y, z)$ with rational coefficients, has infinitely many rational points, then it is birationally equivalent to a line, a conic or a cubic. Mordell's conjecture was first proved by Faltings (1983). Actually Falting's result was not restricted to *plane* algebraic curves, and on the way he proved two other important conjectures of Tate and Shafarevich.

Falting's result implies that the Fermat equation $x^n + y^n = z^n$ has at most finitely many solutions in integers if $n > 3$. In the next section we will see that Wiles' result that semi-stable elliptic curves are modular implies that there are *no* solutions in nonzero integers.

6 Some Applications

The arithmetic of elliptic curves has an interesting application to the ancient problem of congruent numbers. A positive integer n is (confusingly) said to be *congruent* if it is the area of a right-angled triangle whose sides all have rational length, i.e. if there exist positive rational numbers u, v, w such that $u^2 + v^2 = w^2$, $uv = 2n$. For example, 6 is congruent, since it is the area of the right-angled triangle with sides of length 3, 4, 5. Similarly, 5 is congruent, since it is the area of the right-angled triangle with sides of length $3/2, 20/3, 41/6$.

In the margin of his copy of Diophantus' *Arithmetica* Fermat (c. 1640) gave a complete proof that 1 is not congruent. The following is a paraphrase of his argument. Assume that 1 is congruent. Then there exist positive rational numbers u, v, w such that

$$u^2 + v^2 = w^2, \quad uv = 2.$$

Since an integer is a rational square only if it is an integral square, on clearing denominators it follows that there exist positive integers a, b, c, d such that

$$a^2 + b^2 = c^2, \quad 2ab = d^2.$$

Choose such a quadruple a, b, c, d for which c is minimal. Then $(a, b) = 1$. Since d is even, exactly one of a, b is even and we may suppose it to be a. Then

$$a = 2g^2, \quad b = h^2$$

for some positive integers g, h. Since b and c are both odd and $(b, c) = 1$,

$$(c - b, c + b) = 2.$$

Since

$$(c - b)(c + b) = a^2 = 4g^4,$$

it follows that

$$c + b = 2c_1^4, \quad c - b = 2d_1^4,$$

for some relatively prime positive integers c_1, d_1. Then

$$(c_1^2 - d_1^2)(c_1^2 + d_1^2) = c_1^4 - d_1^4 = b = h^2.$$

But

$$(c_1^2 - d_1^2,\ c_1^2 + d_1^2) = 1,$$

since $(c_1^2, d_1^2) = 1$ and b is odd. Hence

$$c_1^2 - d_1^2 = p^2, \quad c_1^2 + d_1^2 = q^2,$$

for some odd positive integers p, q. Thus

$$a_1 = (q + p)/2, \quad b_1 = (q - p)/2$$

are positive integers and

$$a_1^2 + b_1^2 = (q^2 + p^2)/2 = c_1^2,$$
$$2a_1 b_1 = (q^2 - p^2)/2 = d_1^2.$$

Since $c_1 \leq c_1^4 < c$, this contradicts the minimality of c.

It follows that the Fermat equation

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

has no solutions in nonzero integers x, y, z. For if a solution existed and if we put

$$u = 2|yz|/x^2, \quad v = x^2/|yz|, \quad w = (y^4 + z^4)/x^2|yz|,$$

we would have $u^2 + v^2 = w^2, uv = 2$.

It is easily seen that a positive integer n is congruent if and only if there exists a rational number x such that $x, x + n$ and $x - n$ are all rational squares. For suppose

$$x = r^2, \quad x + n = s^2, \quad x - n = t^2,$$

and put

$$u = s - t, \quad v = s + t, \quad w = 2r.$$

Then

$$uv = s^2 - t^2 = 2n$$

and

$$u^2 + v^2 = 2(s^2 + t^2) = 4x = w^2.$$

Conversely, if u, v, w are rational numbers such that $uv = 2n$ and $u^2 + v^2 = w^2$, then

$$(u+v)^2 = w^2 + 4n, \quad (u-v)^2 = w^2 - 4n.$$

Thus, if we put $x = (w/2)^2$, then x, $x+n$ and $x-n$ are all rational squares.

It may be noted that if x is a rational number such that x, $x+n$ and $x-n$ are all rational squares, then $x \neq -n, 0, n$, since $n > 0$ and 2 is not a rational square.

The problem of determining which positive integers are congruent was considered by Arab mathematicians of the 10th century AD, and later by Fibonacci (1225) in his *Liber Quadratorum*. The connection with elliptic curves will now be revealed:

Proposition 13 *A positive integer n is congruent if and only if the cubic curve C_n defined by the polynomial*

$$Y^2 - (X^3 - n^2 X)$$

has a rational point $P = (x, y)$ with $y \neq 0$.

Proof Suppose first that n is congruent. Then there exists a rational number x such that x, $x+n$ and $x-n$ are all rational squares. Hence their product

$$x^3 - n^2 x = x(x-n)(x+n)$$

is also a rational square. Since $x \neq -n, 0, n$, it follows that $x^3 - n^2 x = y^2$, where y is a nonzero rational number.

Suppose now that $P = (x, y)$ is any rational point of the curve C_n with $y \neq 0$. If we put

$$u = |(x^2 - n^2)/y|, \quad v = |2nx/y|, \quad w = |(x^2 + n^2)/y|,$$

then u, v, w are positive rational numbers such that

$$u^2 + v^2 = w^2, \quad uv = 2n. \qquad \square$$

It is readily verified that $\lambda = 1/2$ in the Riemann normal form for C_n.

We now show that, for every positive integer n, the torsion group of C_n has order 4, consisting of the identity element $\boldsymbol{O}$, and the three elements $(0,0)$, $(n,0)$, $(-n,0)$ of order 2. Assume on the contrary that for some positive integer n the curve C_n has a rational point $P = (x, y)$ of finite order with $y \neq 0$ and take n to be the least positive integer with this property. Then $2P = (x', y')$ is also a rational point of C_n of finite order. The formula for the other point of intersection with C_n of the tangent to C_n at P shows that

$$x' = [(x^2 + n^2)/2y]^2.$$

It follows that

$$x' + n = [(x^2 - n^2 + 2nx)/2y]^2,$$
$$x' - n = [(x^2 - n^2 - 2nx)/2y]^2.$$

Moreover x', $x' + n$ and $x' - n$ are all *nonzero* rational squares. Since $2P$ is of finite order, the theorem of Nagell and Lutz mentioned in §5 implies that x' is an integer. Consequently

$$x' = r^2, \quad x' + n = s^2, \quad x' - n = t^2$$

for some positive *integers* r, s, t. Hence n is even, since

$$2n = s^2 - t^2 = (s - t)(s + t)$$

and if one of $s - t$ and $s + t$ is even, so also is the other. Since $n = s^2 - r^2$ and any integral square is congruent to 0 or 1 mod 4, we cannot have $n \equiv 2 \bmod 4$. Hence $n \equiv 0 \bmod 4$. But then $(x'/4, y'/8)$ is a rational point of finite order of $C_{n/4}$, which contradicts the minimality of n.

If n is congruent, then so also is $m^2 n$ for any positive integer m. Thus it is enough to determine which square-free positive integers are congruent. By what we have just proved and Proposition 13, a square-free positive integer n is congruent if and only if the elliptic curve C_n has positive rank. Since C_n admits complex multiplication, a result of Coates and Wiles (1977) shows that if C_n has positive rank, then its L-function vanishes at $s = 1$. (According to the *BSD*-conjecture, C_n has positive rank if and only if its L-function vanishes at $s = 1$.)

By means of the theory of modular forms, Tunnell (1983) has obtained a practical necessary and sufficient condition for the L-function $L(s, C_n)$ of C_n to vanish at $s = 1$: if n is a square-free positive integer, then $L(1, C_n) = 0$ if and only if $A_+(n) = A_-(n)$, where $A_+(n)$, resp. $A_-(n)$, is the number of triples $(x, y, z) \in \mathbb{Z}^3$ with z even, resp. z odd, such that

$$x^2 + 2y^2 + 8z^2 = n \quad \text{if } n \text{ is odd, or } 2x^2 + 2y^2 + 16z^2 = n \quad \text{if } n \text{ is even.}$$

It is not difficult to show that $A_+(n) = A_-(n)$ when $n \equiv 5, 6$ or $7 \bmod 8$, but there seems to be no such simple criterion in other cases. With the aid of a computer it has been verified that, for every $n < 10000$, n is congruent if and only if $A_+(n) = A_-(n)$.

The arithmetic of elliptic curves also has a useful application to the class number problem of Gauss. For any square-free integer $d < 0$, let $h(d)$ be the *class number* of the quadratic field $\mathbb{Q}(\sqrt{d})$. As mentioned in §8 of Chapter IV, it was conjectured by Gauss (1801), and proved by Heilbronn (1934), that $h(d) \to \infty$ as $d \to -\infty$. However, the proof does not provide a method of determining an upper bound for the values of d for which the class number $h(d)$ has a given value. As mentioned in Chapter II, Stark (1967) showed that there are no other negative values of d for which $h(d) = 1$ besides the nine values already known to Gauss. Using methods developed by Baker (1966) for the theory of transcendental numbers, it was shown by Baker (1971) and Stark (1971) that there are exactly 18 negative values of d for which $h(d) = 2$. A simpler and more powerful method for attacking the problem was found by Goldfeld (1976). He obtained an effective lower bound for $h(d)$, provided that there exists a modular elliptic curve over $\mathbb{Q}$ whose L-function has a triple zero at $s = 1$. Gross and Zagier (1986) showed that such an elliptic curve does indeed exist. However, to show that this elliptic curve was modular required a considerable amount of computation. The proof of the *TW*-conjecture makes any computation unnecessary.

The most celebrated application of the arithmetic of elliptic curves has been the recent proof of Fermat's last theorem. In his copy of the translation by Bachet of Diophantus' *Arithmetica* Fermat also wrote "It is impossible to separate a cube into two cubes, or a fourth power into two fourth powers or, in general, any power higher than the second into two like powers. I have discovered a truly marvellous proof of this, which this margin is too narrow to contain."

In other words, Fermat asserted that, if $n > 2$, the equation

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

has no solutions in nonzero integers x, y, z. In §2 of Chapter III we pointed out that it was sufficient to prove his assertion when $n = 4$ and when $n = p$ is an odd prime, and we gave a proof there for $n = 3$.

A nice application to cubic curves of the case $n = 3$ was made by Kronecker (1859). If we make the change of variables

$$x = 2a/(3b - 1), \quad y = (3b + 1)/(3b - 1),$$

with inverse

$$a = x/(y - 1), \quad b = (y + 1)/3(y - 1),$$

then

$$x^3 + y^3 - 1 = 2(4a^3 + 27b^2 + 1)/(3b - 1)^3.$$

Since the equation $x^3 + y^3 = 1$ has no solution in nonzero rational numbers, the only solutions in rational numbers of the equation

$$4a^3 + 27b^2 = -1$$

are $a = -1, b = \pm 1/3$. Consequently the only cubic curves $\mathscr{C}_{a,b}$ with rational coefficients a, b and discriminant $d = -1$ are $Y^2 - X^3 + X \pm 1/3$.

We return now to Fermat's assertion. In the present section we have already given Fermat's own proof for $n = 4$. Suppose now that $p \geq 5$ is prime and assume, contrary to Fermat's assertion, that the equation

$$a^p + b^p + c^p = 0$$

does have a solution in nonzero integers a, b, c. By removing any common factor we may assume that $(a, b) = 1$, and then also $(a, c) = (b, c) = 1$. Since a, b, c cannot all be odd, we may assume that b is even. Then a and c are odd, and we may assume that $a \equiv -1 \bmod 4$.

We now consider the projective cubic curve $\mathscr{E}_{A,B}$ defined by the polynomial

$$Y^2 - X(X - A)(X + B),$$

where $A = a^p$ and $B = b^p$. By construction, $(A, B) = 1$ and

$$A \equiv -1 \bmod 4, \quad B \equiv 0 \bmod 32.$$

Moreover, if we put $C = -(A+B)$, then $C \neq 0$ and $(A, C) = (B, C) = 1$. The linear change of variables

$$X \to 4X, \quad Y \to 8Y + 4X$$

replaces $\mathscr{E}_{A,B}$ by the elliptic curve $\mathscr{W}_{A,B}$ defined by

$$Y^2 + XY - \{X^3 + (B - A - 1)X^2/4 - ABX/16\},$$

which has discriminant

$$\Delta = (ABC)^2/2^8.$$

Our hypotheses ensure that the coefficients of $\mathscr{W}_{A,B}$ are integers and that Δ is a nonzero integer. It may be shown that $\mathscr{W}_{A,B}$ is actually a minimal model for $\mathscr{E}_{A,B}$. Moreover, when we reduce modulo any prime ℓ which divides Δ, the singular point which arises is a node. Thus $\mathscr{W}_{A,B}$ is semi-stable and its conductor N is the product of the distinct primes dividing ABC.

Fermat's last theorem will be proved, for any prime $p \geq 5$, if we show that such an elliptic curve cannot exist if A, B, C are all p-th powers. If p is large, one reason for suspecting that such an elliptic curve cannot exist is that the discriminant is then very large compared with the conductor. Another reason, which does not depend on the size of p, was suggested by Frey (1986). Frey gave a heuristic argument that $\mathscr{W}_{A,B}$ could not then be modular, which would contradict the *TW*-conjecture.

Frey's intuition was made more precise by Serre (1987). Let G be the group of all automorphisms of the field of all algebraic numbers. With any modular form for $\Gamma_0(N)$ one can associate a 2-dimensional representation of G over a finite field. Serre showed that Fermat's last theorem would follow from the *TW*-conjecture, together with a conjecture about lowering the level of such 'Galois representations' associated with modular forms. The latter conjecture was called Serre's ε-conjecture, because it was a special case of a much more general conjecture which Serre made.

Serre's ε-conjecture was proved by Ribet (1990), although the proof might be described as being of order ε^{-1}. Now, for the first time, the falsity of Fermat's last theorem would have a significant consequence: the falsity of the *TW*-conjecture. Since $\mathscr{W}_{A,B}$ is semi-stable with the normalizations made above, to prove Fermat's last theorem it was actually enough to show that any semi-stable elliptic curve was modular. As stated in §5, this was accomplished by Wiles (1995) and Taylor and Wiles (1995). We will not attempt to describe the proof since, besides Fermat's classic excuse, it is beyond the scope of this work.

Fermat's last theorem contributed greatly to the development of mathematics, but Fermat was perhaps lucky that his assertion turned out to be correct. After proving Fermat's assertion for $n = 3$, that the cube of a positive integer could not be the sum of two cubes of positive integers, Euler asserted that, also for any $n \geq 4$, an n-th power of a positive integer could not be expressed as a sum of $n - 1$ n-th powers of positive integers. A counterexample to Euler's conjecture was first found, for $n = 5$, by Lander and Parkin (1966):

$$27^5 + 84^5 + 110^5 + 133^5 = 144^5.$$

Elkies (1988) used the arithmetic of elliptic curves to find infinitely many counterexamples for $n = 4$, the simplest being

$$95800^4 + 217519^4 + 414560^4 = 422481^4.$$

A prize has been offered by Beal (1997) for a proof or disproof of his conjecture that the equation

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

has no solution in coprime positive integers x, y, z if l, m, n are integers > 2. (The exponent 2 must be excluded since, for example, $2^5 + 7^2 = 3^4$ and $2^7 + 17^3 = 71^2$.) Will Beal's conjecture turn out to be like Fermat's or like Euler's?

7 Further Remarks

For sums of squares, see Grosswald [31], Rademacher [46], and Volume II, Chapter IX of Dickson [23]. A recent contribution is Milne [42].

A general reference for the theory of partitions is Andrews [2]. Proposition 4 is often referred to as *Euler's pentagonal number theorem*, since $m(3m-1)/2$ $(m > 1)$ represents the number of dots needed to construct successively larger and larger pentagons. A direct proof of the combinatorial interpretation of Proposition 4 was given by Franklin (1881). It is reproduced in Andrews [2] and in van Lint and Wilson [41]. The replacement of proofs using generating functions by purely combinatorial proofs has become quite an industry; see, for example, Bressoud and Zeilberger [13], [14].

Besides the q-difference equations used in the proof of Proposition 5, there are also *q-integrals*:

$$\int_0^a f(x)d_q x := \sum_{n\geq 0} f(aq^n)(aq^n - aq^{n+1}).$$

The *q-binomial coefficients* (mentioned in §2 of Chapter II)

$$\begin{bmatrix} n \\ m \end{bmatrix} = \begin{bmatrix} n \\ m \end{bmatrix}_q := (q)_n/(q)_m(q)_{n-m} \quad (0 \leq m < n),$$

where $(a)_0 = 1$ and

$$(a)_n = (1-a)(1-aq)\cdots(1-aq^{n-1}) \ (n \geq 1),$$

have recurrence properties similar to those of ordinary binomial coefficients:

$$\begin{bmatrix} n \\ m \end{bmatrix} = \begin{bmatrix} n-1 \\ m-1 \end{bmatrix} + q^m \begin{bmatrix} n-1 \\ m \end{bmatrix} = \begin{bmatrix} n-1 \\ m \end{bmatrix} + q^{n-m} \begin{bmatrix} n-1 \\ m-1 \end{bmatrix} \ (0 < m < n).$$

The *q-hypergeometric series*

$$\sum_{n\geq 0} (a)_n (b)_n x^n/(c)_n (q)_n$$

was already studied by Heine (1847). There is indeed a whole world of q-analysis, which may be regarded as having the same relation to classical analysis as quantum mechanics has to classical mechanics. (The choice of the letter 'q' nearly a century before the advent of quantum mechanics showed remarkable foresight.) There are introductions to this world in Andrews *et al.* [4] and Vilenkin and Klimyk [58]. For Macdonald's conjectures concerning q-analogues of orthogonal polynomials, see Kirillov [36].

Although q-analysis always had its devotees, it remained outside the mainstream of mathematics until recently. Now it arises naturally in the study of *quantum groups*, which are not groups but q-deformations of the universal enveloping algebra of a Lie algebra.

The Rogers–Ramanujan identities were discovered independently by Rogers (1894), Ramanujan (1913) and Schur (1917). Their romantic history is retold in Andrews [2], which contains also generalizations. For the applications of the identities in statistical mechanics, see Baxter's article (pp. 69–84) in Andrews *et al*. [3]. (The same volume contains other interesting articles on mathematical developments arising from Ramanujan's work.)

The Jacobi triple product formula was derived in Chapter XII as the limit of a formula for polynomials. Andrews [1] has given a similar derivation of the Rogers–Ramanujan identities. This approach has found applications and generalizations in conformal field theory, with the two sides of the polynomial identity corresponding to fermionic and bosonic bases for Fock space; see Berkovich and McCoy [9].

These connections go much further than the Rogers–Ramanujan identities. There is now a vast interacting area which involves, besides the theory of partitions, solvable models of statistical mechanics, conformal field theory, integrable systems in classical and quantum mechanics, infinite-dimensional Lie algebras, quantum groups, knot theory and operator algebras. For introductory accounts, see [45], [10] and various articles in [24] and [27]. More detailed treatments of particular aspects are given in Baxter [8], Faddeev and Takhtajan [26], Jantzen [33], Jones [34], Kac [35] and Korepin *et al*. [38].

For the Hardy–Ramanujan–Rademacher expansion for $p(n)$, see Rademacher [46] and Andrews [2]. An interesting proof by means of probability theory for the first term of the expansion has been given by Báez-Duarte [5].

The definition of birational equivalence in §3 is adequate for our purposes, but has been superseded by a more general definition in the language of 'schemes', which is applicable to algebraic varieties of arbitrary dimension without any given embedding in a projective space. For the evolution of the modern concept, see Čižmár [18].

The history of the discovery of the group law on a cubic curve is described by Schappacher [48].

Several good accounts of the arithmetic of elliptic curves are now available; e.g., Knapp [37] and the trilogy [52], [50], [51]. Although the subject has been transformed in the past 25 years, the survey articles by Cassels [16], Tate [55] and Gelbart [28] are still of use. Tate gives a helpful introduction, Cassels has many references to the older literature, and Gelbart explains the connection with the Langlands program, for which see also Gelbart [29].