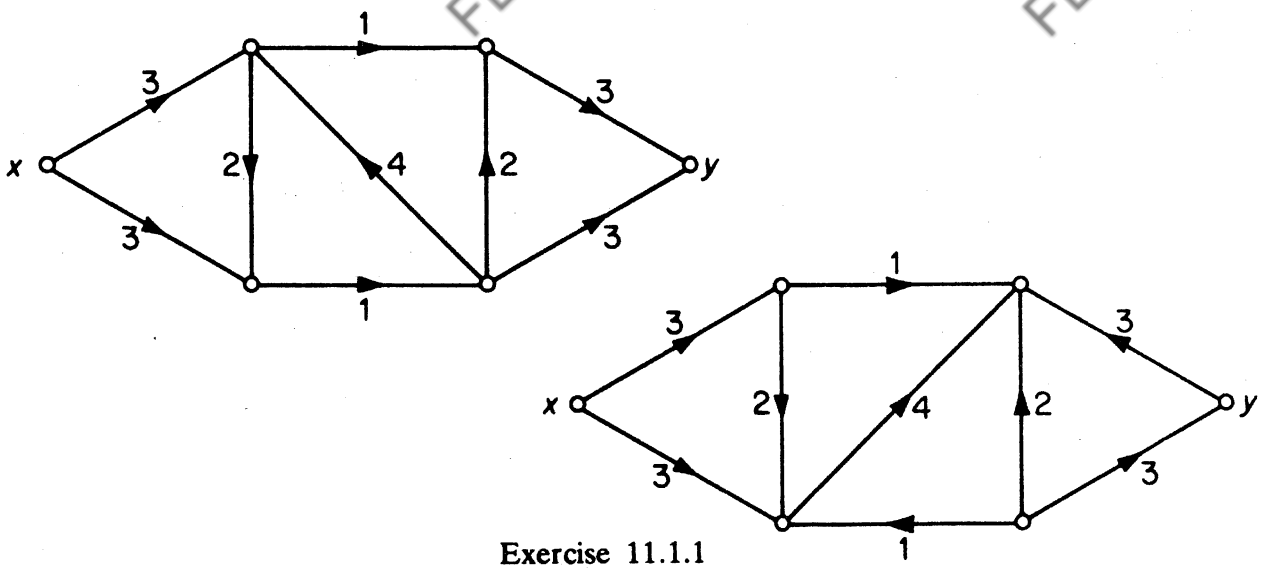


Exercise 11.1.1

11.1.3 Show that, relative to any flow f in N, the resultant flow out of X is equal to the resultant flow into Y.

11.1.4 Show that

(*a*) the function f' given by (11.3) is a flow in N' and that val f' = val f;

(*b*) the restriction to the arc set of N of a flow in N' is a flow in N having the same value.

11.2 CUTS

Let N be a network with a single source x and a single sink y. A *cut* in N is a set of arcs of the form $(S, \bar{S})$, where $x \in S$ and $y \in \bar{S}$. In the network of figure 11.4, a cut is indicated by heavy lines.

The *capacity* of a cut K is the sum of the capacities of its arcs. We denote the capacity of K by cap K; thus

$$\text{cap } K = \sum_{a \in K} c(a)$$

The cut indicated in figure 11.4 has capacity 16.

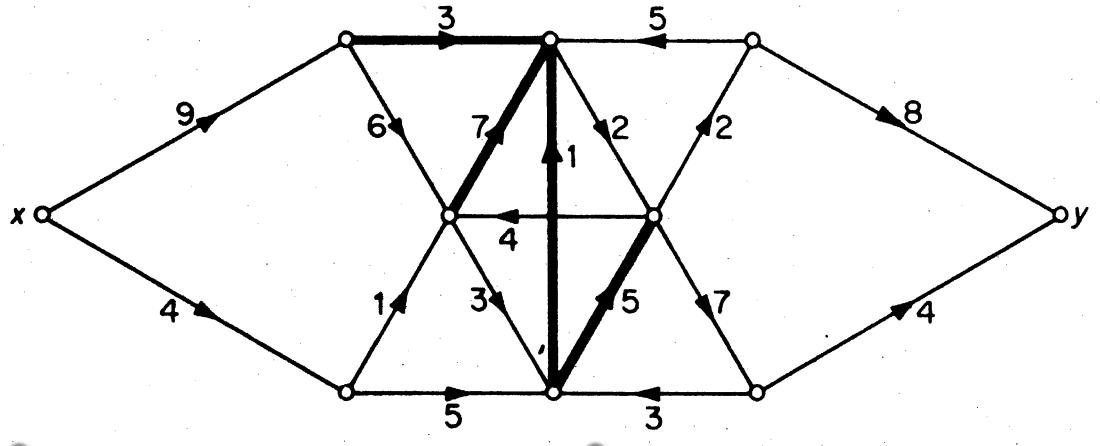


Figure 11.4. A cut in a network

Lemma 11.1 For any flow f and any cut $(S, \bar{S})$ in N

$$\text{val } f = f^+(S) - f^-(S) \tag{11.4}$$

Proof Let f be a flow and $(S, \bar{S})$ a cut in N. From the definitions of flow and value of a flow, we have

$$f^+(v) - f^-(v) = \begin{cases} \text{val } f & \text{if} \quad v = x \\ 0 & \text{if} \quad v \in S \backslash \{x\} \end{cases}$$

Summing these equations over S and simplifying (exercise 11.1.2), we obtain

$$\text{val } f = \sum_{v \in S} (f^+(v) - f^-(v)) = f^+(S) - f^-(S) \quad \square$$

It is convenient to call an arc a *f-zero* if $f(a) = 0$, *f-positive* if $f(a) > 0$, *f-unsaturated* if $f(a) < c(a)$ and *f-saturated* if $f(a) = c(a)$.

Theorem 11.1 For any flow f and any cut $K = (S, \bar{S})$ in N

$$\text{val } f \leq \text{cap } K \tag{11.5}$$

Furthermore, equality holds in (11.5) if and only if each arc in $(S, \bar{S})$ is f-saturated and each arc in $(\bar{S}, S)$ is f-zero.

Proof By (11.1)

$$f^+(S) \leq \text{cap } K \tag{11.6}$$

and

$$f^-(S) \geq 0 \tag{11.7}$$

We obtain (11.5) by substituting inequalities (11.6) and (11.7) in (11.4). The second statement follows, on noting that equality holds in (11.6) if and only if each arc in $(S, \bar{S})$ is f-saturated, and equality holds in (11.7) if and only if each arc in $(\bar{S}, S)$ is f-zero $\square$

A cut K in N is a *minimum cut* if there is no cut K' in N such that $\text{cap } K' < \text{cap } K$. If f^* is a maximum flow and $\tilde{K}$ is a minimum cut, we have, as a special case of theorem 11.1, that

$$\text{val } f^* \leq \text{cap } \tilde{K} \tag{11.8}$$

Corollary 11.1 Let f be a flow and K be a cut such that $\text{val } f = \text{cap } K$. Then f is a maximum flow and K is a minimum cut.

Proof Let f^* be a maximum flow and $\tilde{K}$ a minimum cut. Then, by (11.8),

$$\text{val } f \leq \text{val } f^* \leq \text{cap } \tilde{K} \leq \text{cap } K$$

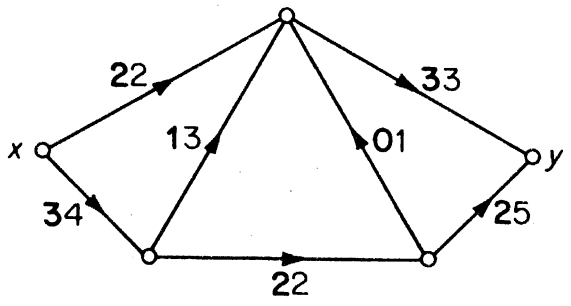
Since, by hypothesis, $\text{val } f = \text{cap } K$, it follows that $\text{val } f = \text{val } f^*$ and $\text{cap } K = \text{cap } \tilde{K}$. Thus f is a maximum flow and K is a minimum cut $\square$

In the next section, we shall prove the converse of corollary 11.1, namely that equality always holds in (11.8).

Exercises

11.2.1 In the following network:

(*a*) determine all cuts;
(*b*) find the capacity of a minimum cut;
(*c*) show that the flow indicated is a maximum flow.



11.2.2 Show that, if there exists no directed (x, y)-path in N, then the value of a maximum flow and the capacity of a minimum cut are both zero.

11.2.3 If $(S, \bar{S})$ and $(T, \bar{T})$ are minimum cuts in N, show that $(S \cup T, \overline{S \cup T})$ and $(S \cap T, \overline{S \cap T})$ are also minimum cuts in N.

11.3 THE MAX-FLOW MIN-CUT THEOREM

In this section we shall present an algorithm for determining a maximum flow in a network. Since a basic requirement of any such algorithm is that it be able to decide when a given flow is, in fact, a maximum flow, we first look at this question.

Let f be a flow in a network N. With each path P in N we associate a non-negative integer $\iota(P)$ defined by

$$\iota(P) = \min_{a \in A(P)} \iota(a)$$

where

$$\iota(a) = \begin{cases} c(a) - f(a) & \text{if } a \text{ is a forward arc of } P \\ f(a) & \text{if } a \text{ is a reverse arc of } P \end{cases}$$

As may easily be seen, $\iota(P)$ is the largest amount by which the flow along P can be increased (relative to f) without violating condition (11.1). The path P is said to be *f-saturated* if $\iota(P) = 0$ and *f-unsaturated* if $\iota(P) > 0$ (or, equivalently, if each forward arc of P is f-unsaturated and each reverse arc of P is f-positive). Put simply, an f-unsaturated path is one that is not being used to its full capacity. An *f-incrementing* path is an f-unsaturated path

from the source x to the sink y. For example, if f is the flow indicated in the network of figure 11.5*a*, then one f-incrementing path is the path $P = xv_1v_2v_3y$. The forward arcs of P are (x, v_1) and (v_3, y) and $\iota(P) = 2$.

The existence of an f-incrementing path P in a network is significant since it implies that f is not a maximum flow; in fact, by sending an additional flow of $\iota(P)$ along P, one obtains a new flow $\hat{f}$ defined by

$$\hat{f}(a) = \begin{cases} f(a) + \iota(P) & \text{if } a \text{ is a forward arc of } P \\ f(a) - \iota(P) & \text{if } a \text{ is a reverse arc of } P \\ f(a) & \text{otherwise} \end{cases} \tag{11.9}$$

for which $\operatorname{val} \hat{f} = \operatorname{val} f + \iota(P)$ (exercise 11.3.1). We shall refer to $\hat{f}$ as the *revised flow based on P*. Figure 11.5*b* shows the revised flow in the network of figure 11.5*a*, based on the f-incrementing path $xv_1v_2v_3y$.

The rôle played by incrementing paths in flow theory is analogous to that of augmenting paths in matching theory, as the following theorem shows (compare theorem 5.1).

Theorem 11.2 A flow f in N is a maximum flow if and only if N contains no f-incrementing path.

Proof If N contains an f-incrementing path P, then f cannot be a maximum flow since $\hat{f}$, the revised flow based on P, has a larger value.

Conversely, suppose that N contains no f-incrementing path. Our aim is to show that f is a maximum flow. Let S denote the set of all vertices to which x is connected by f-unsaturated paths in N. Clearly $x \in S$. Also, since N has no f-incrementing path, $y \in \bar{S}$. Thus $K = (S, \bar{S})$ is a cut in N. We shall show that each arc in $(S, \bar{S})$ is f-saturated and each arc in $(\bar{S}, S)$ is f-zero.

Consider an arc a with tail $u \in S$ and head $v \in \bar{S}$. Since $u \in S$, there exists an f-unsaturated (x, u)-path Q. If a were f-unsaturated, then Q could be extended by the arc a to yield an f-unsaturated (x, v)-path. But $v \in \bar{S}$, and so there is no such path. Therefore a must be f-saturated. Similar reasoning shows that if $a \in (\bar{S}, S)$, then a must be f-zero.

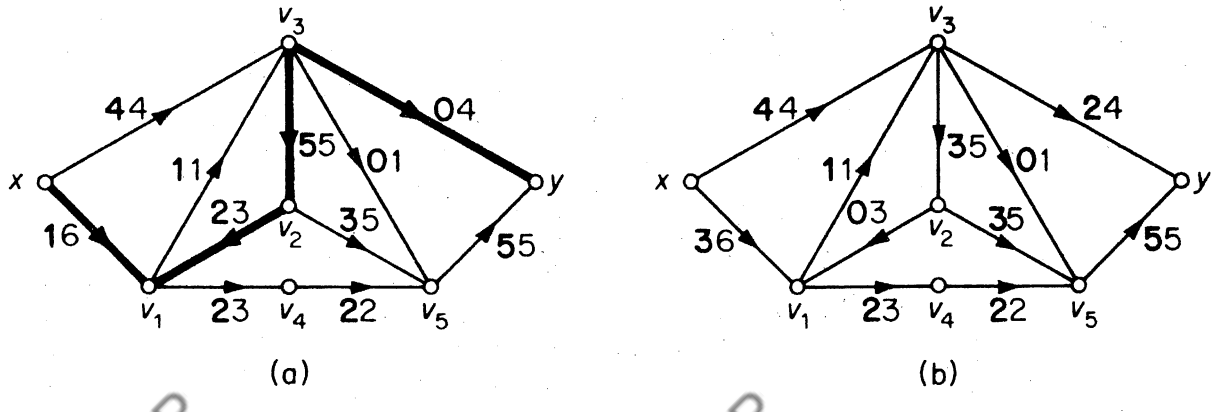


Figure 11.5. (*a*) An f-incrementing path P; (*b*) revised flow based on P

On applying theorem 11.1, we obtain

$$\text{val } f = \text{cap } K$$

It now follows from corollary 11.1 that f is a maximum flow (and that K is a minimum cut) □

In the course of the above proof, we established the existence of a maximum flow f and a minimum cut K such that val f = cap K. We thus have the following theorem, due to Ford and Fulkerson (1956).

Theorem 11.3 In any network, the value of a maximum flow is equal to the capacity of a minimum cut.

Theorem 11.3 is known as the *max-flow min-cut theorem*. It is of central importance in graph theory. Many results on graphs turn out to be easy consequences of this theorem as applied to suitably chosen networks. In sections 11.4 and 11.5 we shall demonstrate two such applications.

The proof of theorem 11.2 is constructive in nature. We extract from it an algorithm for finding a maximum flow in a network. This algorithm, also due to Ford and Fulkerson (1957), is known as the *labelling method*. Starting with a known flow, for instance the zero flow, it recursively constructs a sequence of flows of increasing value, and terminates with a maximum flow. After the construction of each new flow f, a subroutine called the *labelling procedure* is used to find an f-incrementing path, if one exists. If such a path P is found, then $\hat{f}$, the revised flow based on P, is constructed and taken as the next flow in the sequence. If there is no such path, the algorithm terminates; by theorem 11.2, f is a maximum flow.

To describe the labelling procedure we need the following definition. A tree T in N is an *f-unsaturated tree* if (i) $x \in V(T)$, and (ii) for every vertex v of T, the unique (x, v)-path in T is an f-unsaturated path. Such a tree is shown in the network of figure 11.6.

The search for an f-incrementing path involves growing an f-unsaturated tree T in N. Initially, T consists of just the source x. At any stage, there are two ways in which the tree may grow:

1. If there exists an f-unsaturated arc a in $(S, \bar{S})$, where $S = V(T)$, then both a and its head are adjoined to T.

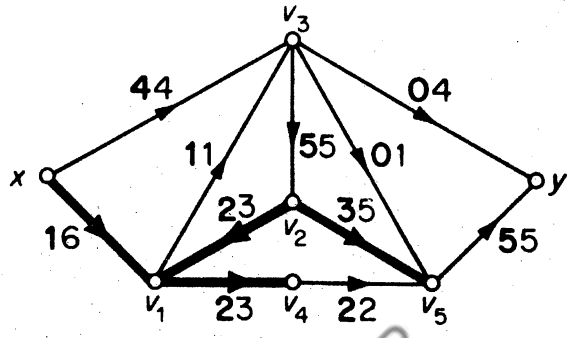


Figure 11.6. An f-unsaturated tree

2. If there exists an f-positive arc a in $(\bar{S}, S)$, then both a and its tail are adjoined to T.

Clearly, each of the above procedures results in an enlarged f-unsaturated tree.

Now either T eventually reaches the sink y or it stops growing before reaching y. The former case is referred to as *breakthrough;* in the event of breakthrough, the (x, y)-path in T is our desired f-incrementing path. If, however, T stops growing before reaching y, we deduce from theorem 11.1 and corollary 11.1 that f is a maximum flow. In figure 11.7, two iterations of this tree-growing procedure are illustrated. The first leads to breakthrough; the second shows that the resulting revised flow is a maximum flow.

The labelling procedure is a systematic way of growing an f-unsaturated tree T. In the process of growing T, it assigns to each vertex v of T the label $l(v) = \iota(P_v)$, where P_v is the unique (x, v)-path in T. The advantage of this labelling is that, in the event of breakthrough, we not only have the f-incrementing path P_y, but also the quantity $\iota(P_y)$ with which to calculate the revised flow based on P_y. The labelling procedure begins by assigning to the source x the label $l(x) = \infty$. It continues according to the following rules:

1. If a is an f-unsaturated arc whose tail u is already labelled but whose head v is not, then v is labelled $l(v) = \min\{l(u), c(a) - f(a)\}$.
2. If a is an f-positive arc whose head u is already labelled but whose tail v is not, then v is labelled $l(v) = \min\{l(u), f(a)\}$.

In each of the above cases, v is said to be labelled *based on* u. To *scan* a labelled vertex u is to label all unlabelled vertices that can be labelled based on u. The labelling procedure is continued until either the sink y is labelled (breakthrough) or all labelled vertices have been scanned and no more vertices can be labelled (implying that f is a maximum flow).

A flow diagram summarising the labelling method is given in figure 11.8.

It is worth pointing out that the labelling method, as described above, is *not* a good algorithm. Consider, for example, the network N in figure 11.9. Clearly, the value of a maximum flow in N is $2m$. The labelling method will use the labelling procedure $2m + 1$ times if it starts with the zero flow and alternates between selecting *xpuvsy* and *xrvuqy* as an incrementing path; for, in each case, the flow value increases by exactly one. Since m is arbitrary, the number of computational steps required to implement the labelling method in this instance can be bounded by no function of ν and ε. In other words, it is not a good algorithm.

However, Edmonds and Karp (1970) have shown that a slight refinement of the labelling procedure turns it into a good algorithm. The refinement suggested by them is the following: in the labelling procedure, scan on a 'first-labelled first-scanned' basis; that is, before scanning a labelled vertex

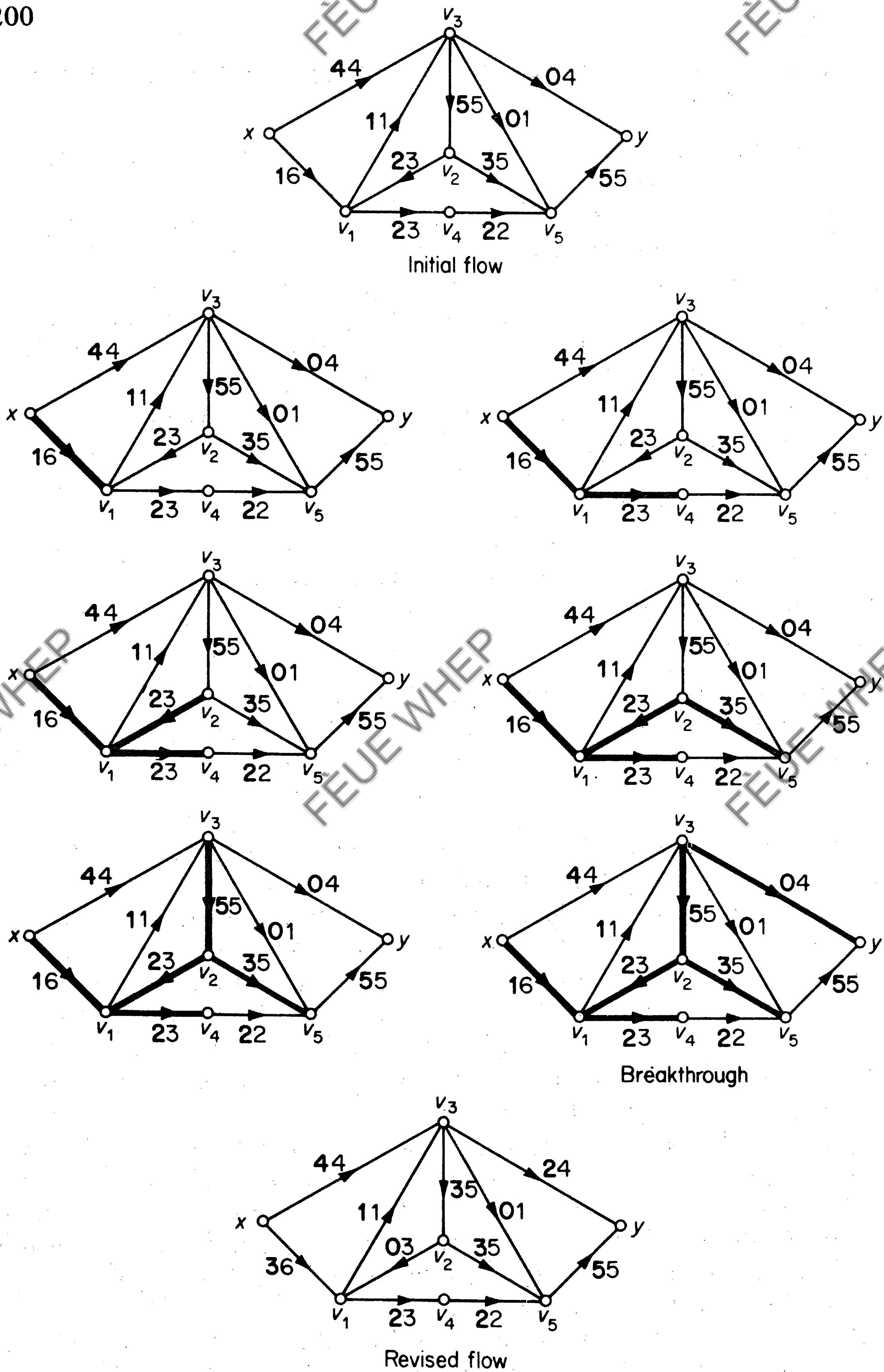


Figure 11.7

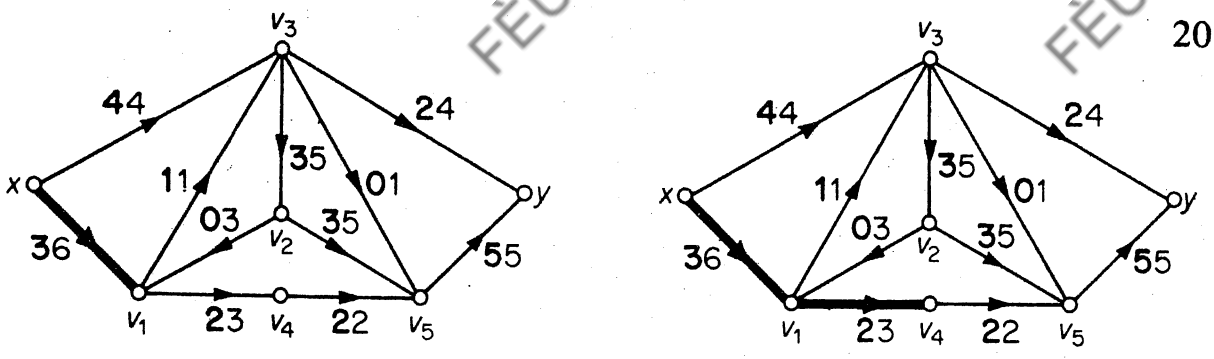


Figure 11.7. (Cont'd)

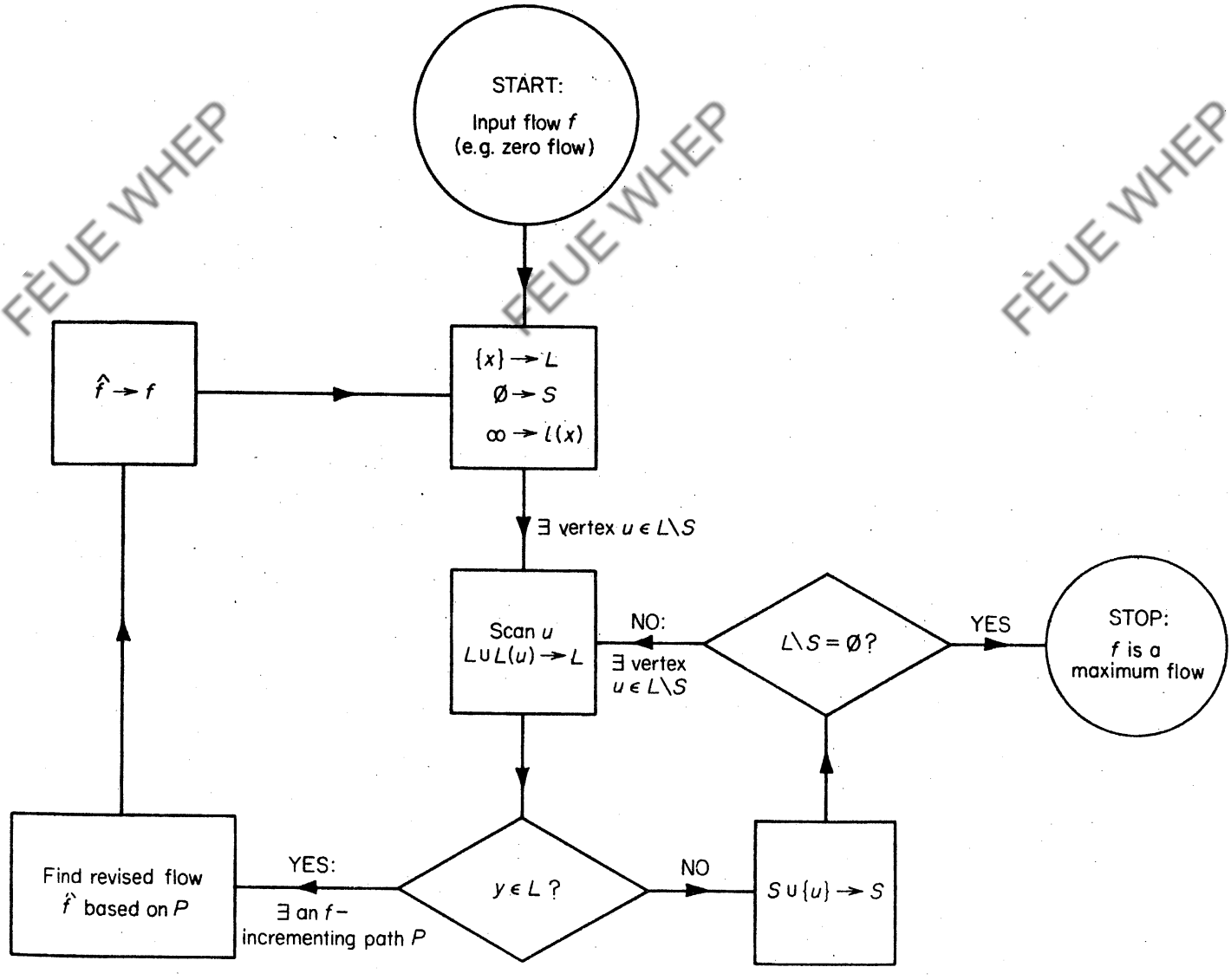


Figure 11.8. The labelling method (L, set of labelled vertices; S, set of scanned vertices; $L(u)$, set of vertices labelled during scanning of u)

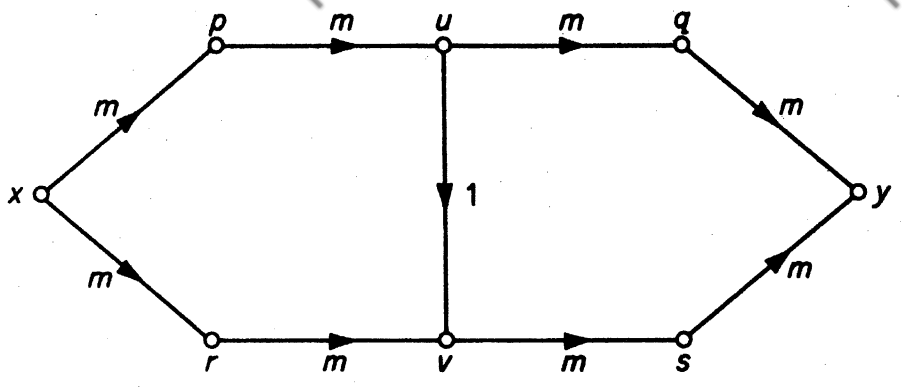


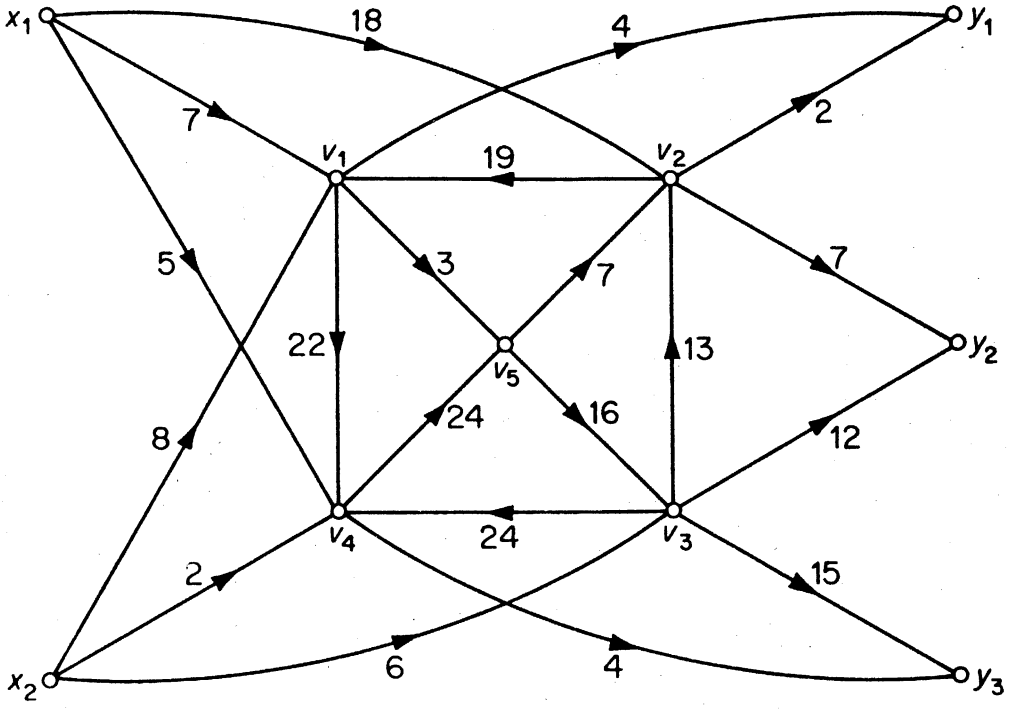
Figure 11.9

u, scan the vertices that were labelled before u. It can be seen that this amounts to selecting a shortest incrementing path. With this refinement, clearly, the maximum flow in the network of figure 11.9 would be found in just two iterations of the labelling procedure.

Exercises

11.3.1 Show that the function $\hat{f}$ given by (11.9) is a flow with $\operatorname{val} \hat{f} = \operatorname{val} f + \iota(P)$.

11.3.2 A certain commodity is produced at two factories x_1 and x_2. The commodity is to be shipped to markets y_1, y_2 and y_3 through the network shown below. Use the labelling method to determine the maximum amount that can be shipped from the factories to the markets.



11.3.3 Show that, in any network N (with integer capacities), there is a maximum flow f such that $f(a)$ is an integer for all $a \in A$.

11.3.4 Consider a network N such that with each arc a is associated an integer $b(a) \leq c(a)$. Modify the labelling method to find a maximum flow f in N subject to the constraint $f(a) \geq b(a)$ for all $a \in A$ (assuming that there is an initial flow satisfying this condition).

11.3.5* Consider a network N such that with each intermediate vertex v is associated a non-negative integer $m(v)$. Show how a maximum flow f satisfying the constraint $f^-(v) \leq m(v)$ for all $v \in V \backslash \{x, y\}$ can be found by applying the labelling method to a modified network.

APPLICATIONS

11.4 MENGER'S THEOREMS

In this section, we shall use the max-flow min-cut theorem to obtain a number of theorems due to Menger (1927); two of these have already been mentioned in section 3.2. The following lemma provides a basic link.

Lemma 11.4 Let N be a network with source x and sink y in which each arc has unit capacity. Then

(a) the value of a maximum flow in N is equal to the maximum number m of arc-disjoint directed (x, y)-paths in N; and
(b) the capacity of a minimum cut in N is equal to the minimum number n of arcs whose deletion destroys all directed (x, y)-paths in N.

Proof Let f^* be a maximum flow in N and let D^* denote the digraph obtained from D by deleting all f^*-zero arcs. Since each arc of N has unit capacity, $f^*(a) = 1$ for all $a \in A(D^*)$. It follows that

(i) $d^+_{D^*}(x) - d^-_{D^*}(x) = \text{val}\, f^* = d^-_{D^*}(y) - d^+_{D^*}(y)$;
(ii) $d^+_{D^*}(v) = d^-_{D^*}(v)$ for all $v \in V \backslash \{x, y\}$.

Therefore (exercise 10.3.3) there exist $\text{val}\, f^*$ arc-disjoint directed (x, y)-paths in D^*, and hence also in D. Thus

$$\text{val}\, f^* \leq m \tag{11.10}$$

Now let $P_1, P_2, \ldots, P_m$ be any system of m arc-disjoint directed (x, y)-paths in N, and define a function f on A by

$$f(a) = \begin{cases} 1 & \text{if } a \text{ is an arc of } \bigcup_{i=1}^{m} P_i \\ 0 & \text{otherwise} \end{cases}$$

Clearly f is a flow in N with value m. Since f^* is a maximum flow, we have

$$\text{val}\, f^* \geq m \tag{11.11}$$

It now follows from (11.10) and (11.11) that

$$\text{val } f^* = m$$

Let $\tilde{K} = (S, \bar{S})$ be a minimum cut in N. Then, in $N - \tilde{K}$, no vertex of $\bar{S}$ is reachable from any vertex in S; in particular, y is not reachable from x. Thus $\tilde{K}$ is a set of arcs whose deletion destroys all directed (x, y)-paths, and we have

$$\text{cap } \tilde{K} = |\tilde{K}| \geq n \tag{11.12}$$

Now let Z be a set of n arcs whose deletion destroys all directed (x, y)-paths, and denote by S the set of all vertices reachable from x in $N - Z$. Since $x \in S$ and $y \in \bar{S}$, $K = (S, \bar{S})$ is a cut in N. Moreover, by the definition of S, $N - Z$ can contain no arc of $(S, \bar{S})$, and so $K \subseteq Z$. Since $\tilde{K}$ is a minimum cut, we conclude that

$$\text{cap } \tilde{K} \leq \text{cap } K = |K| \leq |Z| = n \tag{11.13}$$

Together, (11.12) and (11.13) now yield

$$\text{cap } \tilde{K} = n \quad \square$$

Theorem 11.4 Let x and y be two vertices of a digraph D. Then the maximum number of arc-disjoint directed (x, y)-paths in D is equal to the minimum number of arcs whose deletion destroys all directed (x, y)-paths in D.

Proof We obtain a network N with source x and sink y by assigning unit capacity to each arc of D. The theorem now follows from lemma 11.4 and the max-flow min-cut theorem (11.3) $\square$

A simple trick immediately yields the undirected version of theorem 11.4.

Theorem 11.5 Let x and y be two vertices of a graph G. Then the maximum number of edge-disjoint (x, y)-paths in G is equal to the minimum number of edges whose deletion destroys all (x, y)-paths in G.

Proof Apply theorem 11.4 to $D(G)$, the associated digraph of G (exercise 10.3.6) $\square$

Corollary 11.5 A graph G is k-edge-connected if and only if any two distinct vertices of G are connected by at least k edge-disjoint paths.

Proof This follows directly from theorem 11.5 and the definition of k-edge-connectedness $\square$

We now turn to the vertex versions of the above theorems.

Theorem 11.6 Let x and y be two vertices of a digraph D, such that x is not joined to y. Then the maximum number of internally-disjoint directed (x, y)-paths in D is equal to the minimum number of vertices whose deletion destroys all directed (x, y)-paths in D.

Proof Construct a new digraph D' from D as follows:

(i) split each vertex $v \in V\backslash\{x, y\}$ into two new vertices v' and v'', and join them by an arc (v', v'');
(ii) replace each arc of D with head $v \in V\backslash\{x, y\}$ by a new arc with head v', and each arc of D with tail $v \in V\backslash\{x, y\}$ by a new arc with tail v''. This construction is illustrated in figure 11.10.

Now to each directed (x, y)-path in D' there corresponds a directed (x, y)-path in D obtained by contracting all arcs of type (v', v''); and, conversely, to each directed (x, y)-path in D, there corresponds a directed (x, y)-path in D' obtained by splitting each internal vertex of the path. Furthermore, two directed (x, y)-paths in D' are arc-disjoint if and only if the corresponding paths in D are internally-disjoint. It follows that the maximum number of arc-disjoint directed (x, y)-paths in D' is equal to the maximum number of internally-disjoint directed (x, y)-paths in D. Similarly, the minimum number of arcs in D' whose deletion destroys all directed (x, y)-paths is equal to the minimum number of vertices in D whose deletion destroys all directed (x, y)-paths (exercise 11.4.1). The theorem now follows from theorem 11.4 □

Theorem 11.7 Let x and y be two nonadjacent vertices of a graph G. Then the maximum number of internally-disjoint (x, y)-paths in G is equal to the minimum number of vertices whose deletion destroys all (x, y)-paths.

Proof Apply theorem 11.6 to $D(G)$, the associated digraph of G □

The following corollary is immediate.

Corollary 11.7 A graph G with $\nu \geq k+1$ is k-connected if and only if any two distinct vertices of G are connected by at least k internally-disjoint paths.

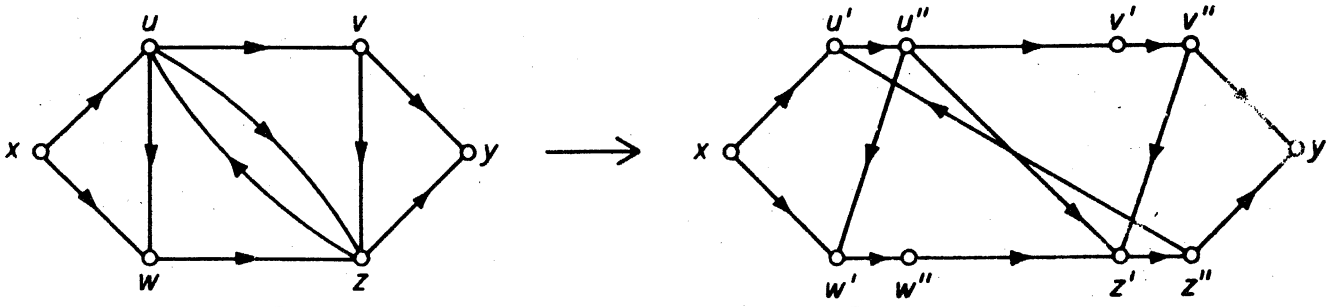


Figure 11.10

Exercises

11.4.1 Show that, in the proof of theorem 11.6, the minimum number of arcs in D' whose deletion destroys all directed (x, y)-paths is equal to the minimum number of vertices in D whose deletion destroys all directed (x, y)-paths.

11.4.2 Derive König's theorem (5.3) from theorem 11.7.

11.4.3 Let G be a graph and let S and T be two disjoint subsets of V. Show that the maximum number of vertex-disjoint paths with one end in S and one end in T is equal to the minimum number of vertices whose deletion separates S from T (that is, after deletion no component contains a vertex of S and a vertex of T).

11.4.4* Show that if G is k-connected with $k \geq 2$, then any k vertices of G are contained together in some cycle. (G. A. Dirac)

11.5 FEASIBLE FLOWS

Let N be a network. Suppose that to each source x_i of N is assigned a non-negative integer $\sigma(x_i)$, called the *supply* at x_i, and to each sink y_j of N is assigned a non-negative integer $\partial(y_j)$, called the *demand* at y_j. A flow f in N is said to be *feasible* if

$$f^+(x_i) - f^-(x_i) \leq \sigma(x_i) \quad \text{for all} \quad x_i \in X$$

and

$$f^-(y_j) - f^+(y_j) \geq \partial(y_j) \quad \text{for all} \quad y_j \in Y$$

In other words, a flow f is feasible if the resultant flow out of each source x_i relative to f does not exceed the supply at x_i, and the resultant flow into each sink y_j relative to f is at least as large as the demand at y_j. A natural question, then, is to ask for necessary and sufficient conditions for the existence of a feasible flow in N. Theorem 11.8, due to Gale (1957), provides an answer to this question. It says that a feasible flow exists if and only if, for every subset S of V, the total capacity of arcs from S to $\bar{S}$ is at least as large as the net demand of $\bar{S}$.

For any subset S of V, we shall denote $\sum_{v \in S} \sigma(v)$ by $\sigma(S)$ and $\sum_{v \in S} \partial(v)$ by $\partial(S)$.

Theorem 11.8 There exists a feasible flow in N if and only if, for all $S \subseteq V$

$$c(S, \bar{S}) \geq \partial(Y \cap \bar{S}) - \sigma(X \cap \bar{S}) \tag{11.14}$$

Proof Construct a new network N' from N as follows:

(i) adjoin two new vertices x and y to N;
(ii) join x to each $x_i \in X$ by an arc of capacity $\sigma(x_i)$;

(iii) join each $y_j \in Y$ to y by an arc of capacity $\partial(y_j)$;
(iv) designate x as the source and y as the sink of N'.

This construction is illustrated in figure 11.11.

It is not difficult to see that N has a feasible flow if and only if N' has a flow that saturates each arc of the cut $(Y, \{y\})$ (exercise 11.5.1). Now a flow in N' that saturates each arc of $(Y, \{y\})$ clearly has value $\partial(Y) = \text{cap}(Y, \{y\})$, and is therefore, by corollary 11.1, a maximum flow. It follows that N has a feasible flow if and only if, for each cut $(S \cup \{x\}, \bar{S} \cup \{y\})$ of N'

$$\text{cap}(S \cup \{x\}, \bar{S} \cup \{y\}) \geq \partial(Y) \qquad (11.15)$$

But conditions (11.14) and (11.15) are precisely the same; for, denoting the capacity function in N' by c', we have

$$\begin{aligned} \text{cap}(S \cup \{x\}, \bar{S} \cup \{y\}) &= c'(S, \bar{S}) + c'(S, \{y\}) + c'(\{x\}, \bar{S}) \\ &= c(S, \bar{S}) + \partial(Y \cap S) + \sigma(X \cap \bar{S}) \quad \square \end{aligned}$$

There are many applications of theorem 11.8 to problems in graph theory. We shall discuss one such application.

Let $\mathbf{p} = (p_1, p_2, \ldots, p_m)$ and $\mathbf{q} = (q_1, q_2, \ldots, q_n)$ be two sequences of non-negative integers. We say that the pair $(\mathbf{p}, \mathbf{q})$ is *realisable by a simple bipartite graph* if there exists a simple bipartite graph G with bipartition $(\{x_1, x_2, \ldots, x_m\}, \{y_1, y_2, \ldots, y_n\})$, such that

$$d(x_i) = p_i \quad \text{for} \quad 1 \leq i \leq m$$

and

$$d(y_j) = q_j \quad \text{for} \quad 1 \leq j \leq n$$

For example, the pair $(\mathbf{p}, \mathbf{q})$, where

$$\mathbf{p} = (3, 2, 2, 2, 1) \quad \text{and} \quad \mathbf{q} = (3, 3, 2, 1, 1)$$

is realisable by the bipartite graph of figure 11.12.

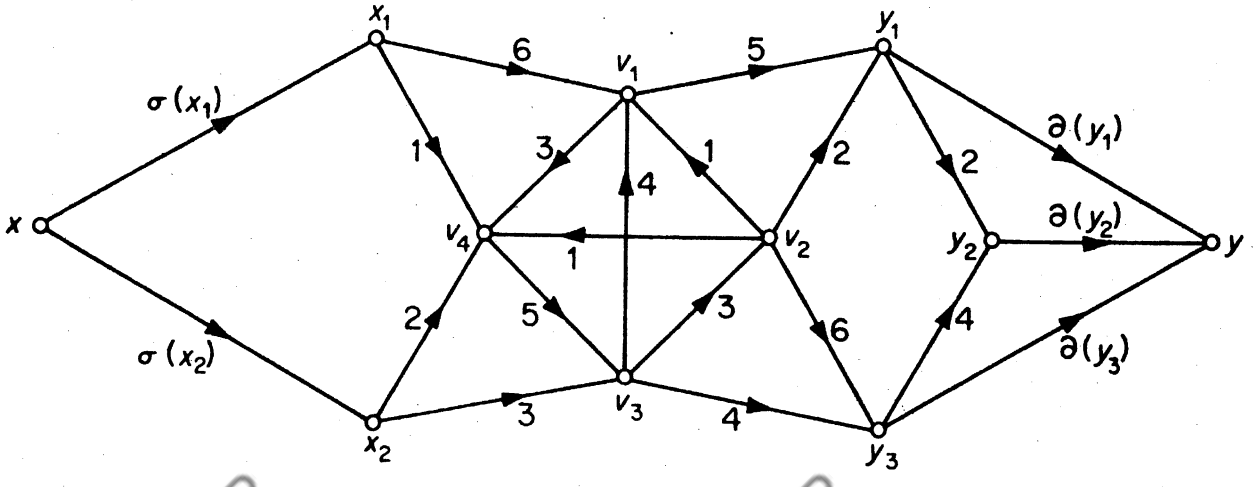


Figure 11.11

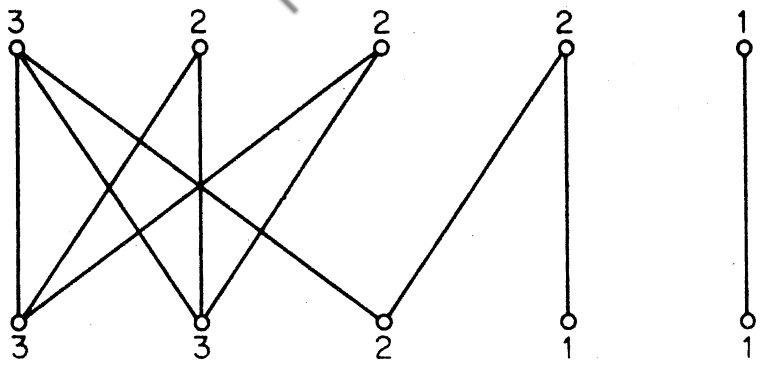


Figure 11.12

An obvious necessary condition for realisability is that

$$\sum_{i=1}^{m} p_i = \sum_{j=1}^{n} q_j \tag{11.16}$$

However, (11.16) is not in itself sufficient. For instance, the pair $(\mathbf{p}, \mathbf{q})$, where

$$\mathbf{p} = (5, 4, 4, 2, 1) \quad \text{and} \quad \mathbf{q} = (5, 4, 4, 2, 1)$$

is not realisable by any simple bipartite graph (exercise 11.5.2). In the following theorem we present necessary and sufficient conditions for the realisability of a pair of sequences by a simple bipartite graph. The order of the terms in the sequences clearly has no bearing on the question of realisability, and we shall find it convenient to assume that the terms of $\mathbf{q}$ are arranged in nonincreasing order

$$q_1 \geq q_2 \geq \ldots \geq q_n \tag{11.17}$$

Theorem 11.9 Let $\mathbf{p} = (p_1, p_2, \ldots, p_m)$ and $\mathbf{q} = (q_1, q_2, \ldots, q_n)$ be two sequences of non-negative integers that satisfy (11.16) and (11.17). Then $(\mathbf{p}, \mathbf{q})$ is realisable by a simple bipartite graph if and only if

$$\sum_{i=1}^{m} \min\{p_i, k\} \geq \sum_{j=1}^{k} q_j \quad \text{for} \quad 1 \leq k \leq n \tag{11.18}$$

Proof Let $X = \{x_1, x_2, \ldots, x_m\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$ be two disjoint sets, and let D be the digraph obtained from the complete bipartite graph with bipartition (X, Y) by orienting each edge from X to Y. We obtain a network N by assigning unit capacity to each arc of D and designating the vertices in X and Y as its sources and sinks, respectively. We shall assume, further, that the supply at source x_i is p_i, $1 \leq i \leq m$, and that the demand at sink y_j is q_j, $1 \leq j \leq n$.

Now, to each spanning subgraph of D, there corresponds a flow in N which saturates precisely the arcs of the subgraph, and this correspondence is clearly one-one. In view of (11.16), it follows that $(\mathbf{p}, \mathbf{q})$ is realisable by a

simple bipartite graph if and only if the network N has a feasible flow. We now use theorem 11.8.

For any set S of vertices in N, write

$$I(S)=\{i \mid x_i \in S\} \quad \text{and} \quad J(S)=\{j \mid y_j \in S\}$$

Then, by definition,

$$\left.\begin{aligned} c(S, \bar{S}) &= |I(S)|\,|J(\bar{S})| \\ \sigma(X \cap \bar{S}) = \sum_{i \in I(\bar{S})} p_i \quad &\text{and} \quad \partial(Y \cap \bar{S}) = \sum_{j \in J(\bar{S})} q_j \end{aligned}\right\} \tag{11.19}$$

Suppose that N has a feasible flow. By theorem 11.8 and (11.19)

$$|I(S)|\,|J(\bar{S})| \geq \sum_{j \in J(\bar{S})} q_j - \sum_{i \in I(\bar{S})} p_i$$

for any $S \subseteq X \cup Y$. Setting $S=\{x_i \mid p_i > k\} \cup \{y_j \mid j > k\}$, we have

$$\sum_{i \in I(S)} \min\{p_i, k\} \geq \sum_{j=1}^{k} q_j - \sum_{i \in I(\bar{S})} \min\{p_i, k\}$$

Since this holds for all values of k, (11.18) follows.

Conversely, suppose that (11.18) is satisfied. Let S be any set of vertices in N. By (11.18) and (11.19)

$$c(S, \bar{S}) \geq \sum_{i \in I(S)} \min\{p_i, k\} \geq \sum_{j=1}^{k} q_j - \sum_{j \in I(\bar{S})} \min\{p_i, k\} \geq \partial(Y \cap \bar{S}) - \sigma(X \cap \bar{S})$$

where $k=|J(\bar{S})|$. It follows from theorem 11.8 that N has a feasible flow □

We conclude by looking at theorem 11.9 from the viewpoint of matrices. With each simple bipartite graph G having bipartition $(\{x_1, x_2, \ldots, x_m\}, \{y_1, y_2, \ldots, y_n\})$, we can associate an $m \times n$ matrix $\mathbf{B}$ in which $b_{ij}=1$ or 0, depending on whether $x_i y_j$ is an edge of G or not. Conversely, every $m \times n$ (0, 1)-matrix corresponds in this way to a simple bipartite graph. Thus theorem 11.9 provides necessary and sufficient conditions for the existence of an $m \times n$ (0, 1)-matrix $\mathbf{B}$ with row sums $p_1, p_2, \ldots, p_m$ and column sums $q_1, q_2, \ldots, q_n$.

There is a simple way of visualising condition (11.18) in terms of matrices. Let $\mathbf{B}^*$ denote the (0, 1)-matrix in which the p_i leading terms in each row i are ones, and the remaining entries are zeros, and let $p_1^*, p_2^*, \ldots, p_n^*$ be the column sums of $\mathbf{B}^*$. The sequence $\mathbf{p}^*=(p_1^*, p_2^*, \ldots, p_n^*)$ is called the *conjugate* of $\mathbf{p}$. The conjugate of (5, 4, 4, 2, 1) is (5, 4, 3, 3, 1), for example (see figure 11.13).

Now consider the sum $\sum_{j=1}^{k} p_j^*$. Row i of $\mathbf{B}^*$ contributes $\min\{p_i, k\}$ to this sum. Therefore the left-hand side of (11.18) is equal to $\sum_{j=1}^{k} p_j^*$, and (11.18) is

$$\begin{array}{cc} & \mathbf{p}^* \\ & \begin{array}{ccccc} 5 & 4 & 3 & 3 & 1 \end{array} \\ \mathbf{p} \begin{array}{c} 5 \\ 4 \\ 4 \\ 2 \\ 1 \end{array} & \left[\begin{array}{ccccc} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{array}\right] \end{array}$$

Figure 11.13

equivalent to the condition

$$\sum_{j=1}^{k} p_j^* \geq \sum_{j=1}^{k} q_j \quad \text{for} \quad 1 \leq k \leq n$$

This formulation of theorem 11.9 in terms of (0, 1)-matrices is due to Ryser (1957). For other applications of the theory of flows in networks, we refer the reader to Ford and Fulkerson (1962).

Exercises

11.5.1 Show that the network N in the proof of theorem 11.8 has a feasible flow if and only if N' has a flow that saturates each arc of the cut $(Y, \{y\})$.

11.5.2 Show that the pair $(\mathbf{p}, \mathbf{q})$, where

$$\mathbf{p} = (5, 4, 4, 2, 1) \quad \text{and} \quad \mathbf{q} = (5, 4, 4, 2, 1)$$

is not realisable by any simple bipartite graph.

11.5.3 Given two sequences, $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ and $\mathbf{q} = (q_1, q_2, \ldots, q_n)$, find necessary and sufficient conditions for the existence of a digraph D on the vertex set $\{v_1, v_2, \ldots, v_n\}$, such that (i) $d^-(v_i) = p_i$ and $d^+(v_i) = q_i$, $1 \leq i \leq n$, and (ii) D has a (0, 1) adjacency matrix.

11.5.4* Let $\mathbf{p} = (p_1, p_2, \ldots, p_m)$ and $\mathbf{q} = (q_1, q_2, \ldots, q_n)$ be two nonincreasing sequences of non-negative integers, and denote the sequences $(p_2, p_3, \ldots, p_m)$ and $(q_1 - 1, q_2 - 1, \ldots, q_{p_1} - 1, q_{p_1+1}, \ldots, q_n)$ by $\mathbf{p}'$ and $\mathbf{q}'$, respectively.

(*a*) Show that $(\mathbf{p}, \mathbf{q})$ is realisable by a simple bipartite graph if and only if the same is true of $(\mathbf{p}', \mathbf{q}')$.

(*b*) Using (*a*), describe an algorithm for constructing a simple bipartite graph which realises $(\mathbf{p}, \mathbf{q})$, if such a realisation exists.

11.5.5 An $(m+n)$-regular graph G is (m,n)-*orientable* if it can be oriented so that each indegree is either m or n.

(*a*)* Show that G is (m,n)-orientable if and only if there is a partition (V_1, V_2) of V such that, for every $S \subseteq V$,

$$|(m-n)(|V_1 \cap S| - |V_2 \cap S|)| \leq |[S, \bar{S}]|$$

(*b*) Deduce that if G is (m,n)-orientable and $m > n$, then G is also $(m-1, n+1)$-orientable.

REFERENCES

Edmonds, J. and Karp, R. M. (1972). Theoretical improvements in algorithmic efficiency for network flow problems. *J. Assoc. Comput. Mach.*, **19,** 248–64

Ford, L. R. Jr. and Fulkerson, D. R. (1956). Maximal flow through a network. *Canad. J. Math.*, **8,** 399–404

Ford, L. R. Jr. and Fulkerson, D. R. (1957). A simple algorithm for finding maximal network flows and an application to the Hitchcock problem. *Canad. J. Math.*, **9,** 210–18

Ford, L. R. Jr. and Fulkerson, D. R. (1962). *Flows in Networks*, Princeton University Press, Princeton

Gale, D. (1957). A theorem on flows in networks. *Pacific J. Math.*, **7,** 1073–82

Menger, K. (1927). Zur allgemeinen Kurventheorie. *Fund. Math.*, **10,** 96–115

Ryser, H. J. (1957). Combinatorial properties of matrices of zeros and ones. *Canad. J. Math.*, **9,** 371–77

12 The Cycle Space and Bond Space

12.1 CIRCULATIONS AND POTENTIAL DIFFERENCES

Let D be a digraph. A real-valued function f on A is called a *circulation* in D if it satisfies the conservation condition at each vertex:

$$f^-(v) = f^+(v) \quad \text{for all} \quad v \in V \tag{12.1}$$

If we think of D as an electrical network, then such a function f represents a circulation of currents in D. Figure 12.1 shows a circulation in a digraph.

If f and g are any two circulations and r is any real number, then it is easy to verify that both $f+g$ and rf are also circulations. Thus the set of all circulations in D is a vector space. We denote this space by $\mathscr{C}$. In what follows, we shall find it convenient to identify a subset S of A with $D[S]$, the subdigraph of D induced by S.

There are certain circulations of special interest. These are associated with cycles in D. Let C be a cycle in D with an assigned orientation and let C^+ denote the set of arcs of C whose direction agrees with this orientation. We associate with C the function f_C defined by

$$f_C(a) = \begin{cases} 1 & \text{if} \quad a \in C^+ \\ -1 & \text{if} \quad a \in C \backslash C^+ \\ 0 & \text{if} \quad a \notin C \end{cases}$$

Clearly, f_C satisfies (12.1) and hence is a circulation. Figure 12.2 depicts a circulation associated with a cycle.

We shall see later on that each circulation is a linear combination of the circulations associated with cycles. For this reason we refer to $\mathscr{C}$ as the *cycle space* of D.

We now turn our attention to a related class of functions. Given a function p on the vertex set V of D, we define the function δp on the arc set A by the rule that, if an arc a has tail x and head y, then

$$\delta p(a) = p(x) - p(y) \tag{12.2}$$

If D is thought of as an electrical network with potential $p(v)$ at v, then, by (12.2), δp represents the potential difference along the wires of the network. For this reason a function g on A is called a *potential difference* in D if

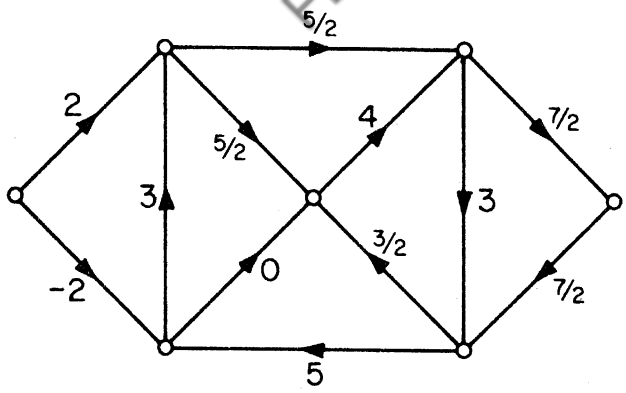


Figure 12.1. A circulation

$g = \delta p$ for some function p on V. Figure 12.3 shows a digraph with an assignment of potentials to its vertices and the corresponding potential difference.

As with circulations, the set $\mathcal{B}$ of all potential differences in D is closed under addition and scalar multiplication and, hence, is a vector space.

Analogous to the function f_C associated with a cycle C, there is a function g_B associated with a bond B. Let $B = [S, \bar{S}]$ be a bond of D. We define g_B by

$$g_B(a) = \begin{cases} 1 & \text{if} \quad a \in (S, \bar{S}) \\ -1 & \text{if} \quad a \in (\bar{S}, S) \\ 0 & \text{if} \quad a \notin B \end{cases}$$

It can be verified that $g_B = \delta p$ where

$$p(v) = \begin{cases} 1 & \text{if} \quad v \in S \\ 0 & \text{if} \quad v \in \bar{S} \end{cases}$$

Figure 12.4 depicts the potential difference associated with a bond.

We shall see that each potential difference is a linear combination of potential differences associated with bonds. For this reason we refer to $\mathcal{B}$ as the *bond space* of D.

In studying the properties of the two vector spaces $\mathcal{B}$ and $\mathcal{C}$, we shall find

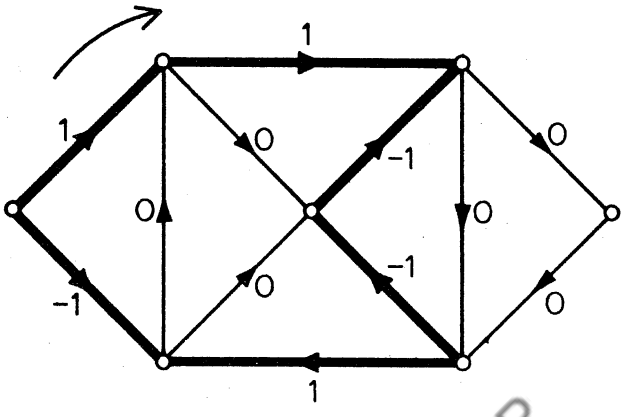


Figure 12.2

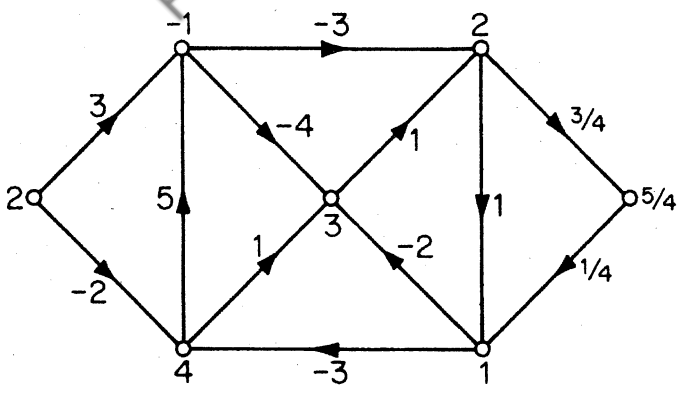


Figure 12.3. A potential difference

it convenient to regard a function on A as a row vector whose coordinates are labelled with the elements of A. The relationship between $\mathscr{B}$ and $\mathscr{C}$ is best seen by introducing the incidence matrix of D. With each vertex v of D we associate the function m_v on A defined by

$$m_v(a) = \begin{cases} 1 & \text{if } a \text{ is a link and } v \text{ is the tail of } a \\ -1 & \text{if } a \text{ is a link and } v \text{ is the head of } a \\ 0 & \text{otherwise} \end{cases}$$

The *incidence matrix* of D is the matrix $\boldsymbol{M}$ whose rows are the functions m_v. Figure 12.5 shows a digraph and its incidence matrix.

Theorem 12.1 Let $\boldsymbol{M}$ be the incidence matrix of a digraph D. Then $\mathscr{B}$ is the row space of $\boldsymbol{M}$ and $\mathscr{C}$ is its orthogonal complement.

Proof Let $g = \delta p$ be a potential difference in D. It follows from (12.2) that

$$g(a) = \sum_{v \in V} p(v) m_v(a) \quad \text{for all} \quad a \in A$$

Thus g is a linear combination of the rows of $\boldsymbol{M}$. Conversely, any linear combination of the rows of $\boldsymbol{M}$ is a potential difference. Hence $\mathscr{B}$ is the row space of $\boldsymbol{M}$.

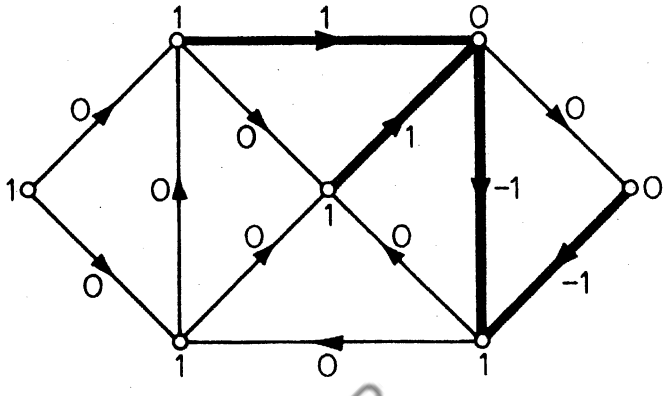
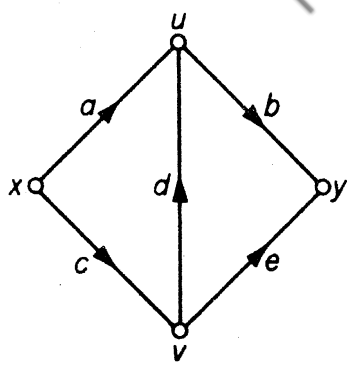


Figure 12.4



(a)

	a	b	c	d	e
x	1	0	1	0	0
u	-1	1	0	-1	0
v	0	0	-1	1	1
y	0	-1	0	0	-1

(b)

Figure 12.5. (*a*) D; (*b*) the incidence matrix of D

Now let f be a function on A. The condition (12.1) for f to be a circulation can be rewritten as

$$\sum_{a \in A} m_v(a) f(a) = 0 \quad \text{for all} \quad v \in V$$

This implies that f is a circulation if and only if it is orthogonal to each row of $\boldsymbol{M}$. Hence $\mathscr{C}$ is the orthogonal complement of $\mathscr{B}$ □

The *support* of a function f on A is the set of elements of A at which the value of f is nonzero. We denote the support of f by $\|f\|$.

Lemma 12.2.1 If f is a nonzero circulation, then $\|f\|$ contains a cycle.

Proof This follows immediately, since $\|f\|$ clearly cannot contain a vertex of degree one □

Lemma 12.2.2 If g is a nonzero potential difference, then $\|g\|$ contains a bond.

Proof Let $g = \delta p$ be a nonzero potential difference in D. Choose a vertex $u \in V$ which is incident with an arc of $\|g\|$ and set

$$U = \{v \in V \mid p(v) = p(u)\}$$

Clearly, $\|g\| \supseteq [U, \bar{U}]$ since $g(a) \neq 0$ for all $a \in [U, \bar{U}]$. But, by the choice of u, $[U, \bar{U}]$ is nonempty. Thus $\|g\|$ contains a bond □

A matrix $\mathbf{B}$ is called a *basis matrix* of $\mathscr{B}$ if the rows of $\mathbf{B}$ form a basis for $\mathscr{B}$; a basis matrix of $\mathscr{C}$ is similarly defined. We shall find the following notation convenient. If $\mathbf{R}$ is a matrix whose columns are labelled with the elements of A, and if $S \subseteq A$, we shall denote by $\mathbf{R} \mid S$ the submatrix of $\mathbf{R}$ consisting of those columns of $\mathbf{R}$ labelled with elements in S. If $\mathbf{R}$ has a single row, our notation is the same as the usual notation for the restriction of a function to a subset of its domain.

Theorem 12.2 Let **B** and **C** be basis matrices of $\mathscr{B}$ and $\mathscr{C}$, respectively. Then, for any $S \subseteq A$

(i) the columns of $\mathbf{B} \mid S$ are linearly independent if and only if S is acyclic, and
(ii) the columns of $\mathbf{C} \mid S$ are linearly independent if and only if S contains no bond.

Proof Denote the column of **B** corresponding to arc a by $\mathbf{B}(a)$. The columns of $\mathbf{B} \mid S$ are linearly dependent if and only if there exists a function f on A such that

$$f(a) \neq 0 \quad \text{for some} \quad a \in S$$
$$f(a) = 0 \quad \text{for all} \quad a \notin S$$

and

$$\sum_{a \in A} f(a)\mathbf{B}(a) = \mathbf{O}$$

We conclude that the columns of $\mathbf{B} \mid S$ are linearly dependent if and only if there exists a nonzero circulation f such that $\|f\| \subseteq S$. Now if there is such an f then, by lemma 12.2.1, S contains a cycle. On the other hand, if S contains a cycle C, then f_C is a nonzero circulation with $\|f_C\| = C \subseteq S$. It follows that the columns of $\mathbf{B} \mid S$ are linearly independent if and only if S is acyclic. A similar argument using lemma 12.2.2 yields a proof of (ii) □

Corollary 12.2 The dimensions of $\mathscr{B}$ and $\mathscr{C}$ are given by

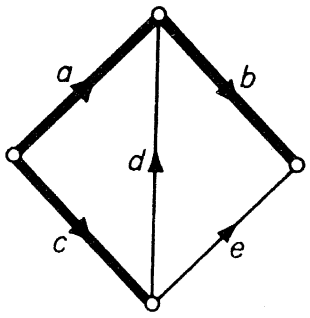
$$\dim \mathscr{B} = \nu - \omega \tag{12.3}$$
$$\dim \mathscr{C} = \varepsilon - \nu + \omega \tag{12.4}$$

Proof Consider a basis matrix **B** of $\mathscr{B}$. By theorem 12.2

$$\text{rank } \mathbf{B} = \max\{|S| \mid S \subseteq A, S \text{ acyclic}\}$$

The above maximum is attained when S is a maximal forest of D, and is therefore (exercise 2.2.4) equal to $\nu - \omega$. Since $\dim \mathscr{B} = \text{rank } \mathbf{B}$, this establishes (12.3). Now (12.4) follows, since $\mathscr{C}$ is the orthogonal complement of $\mathscr{B}$ □

Let T be a maximal forest of D. Associated with T is a special basis matrix of $\mathscr{C}$. If a is an arc of $\bar{T}$, then $T + a$ contains a unique cycle. Let C_a denote this cycle and let f_a denote the circulation corresponding to C_a, defined so that $f_a(a) = 1$. The $(\varepsilon - \nu + \omega) \times \varepsilon$ matrix **C** whose rows are f_a, $a \in \bar{T}$, is a basis matrix of $\mathscr{C}$. This follows from the fact that each row is a circulation and that rank $\mathbf{C} = \varepsilon - \nu + \omega$ (because $\mathbf{C} \mid \bar{T}$ is an identity matrix). We refer to **C** as the basis matrix of $\mathscr{C}$ *corresponding to* T. Figure 12.6*b* shows the basis matrix of $\mathscr{C}$ corresponding to the tree indicated in figure 12.6*a*.



(a)

	a	b	c	d	e
f_d	−1	0	1	1	0
f_e	−1	−1	1	0	1

(b)

	a	b	c	d	e
g_a	1	0	0	1	1
g_b	0	1	0	0	1
g_c	0	0	1	−1	−1

(c)

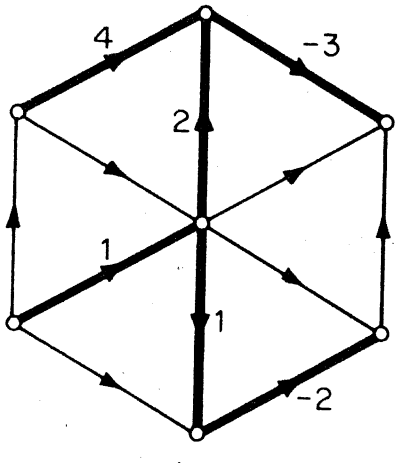
Figure 12.6

Analogously, if a is an arc of T, then $\bar{T}+a$ contains a unique bond (see theorem 2.6). Let B_a denote this bond and g_a the potential difference corresponding to B_a, defined so that $g_a(a)=1$. The $(\nu-\omega)\times\varepsilon$ matrix **B** whose rows are g_a, $a\in T$, is a basis matrix of $\mathscr{B}$, called the basis matrix of $\mathscr{B}$ *corresponding to* T. Figure 12.6c gives an example of such a matrix.

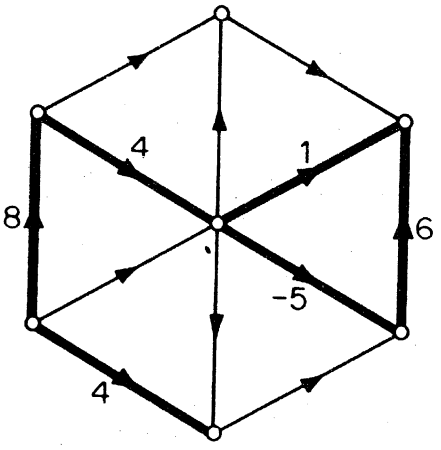
The relationship between cycles and bonds that has become apparent from the foregoing discussion finds its proper setting in the theory of matroids. The interested reader is referred to Tutte (1971).

Exercises

12.1.1 (*a*) In figure (i) below is indicated a function on a spanning tree and in figure (ii) a function on the complement of the tree. Extend the function in (i) to a potential difference and the function in (ii) to a circulation.



(i)



(ii)

(*b*) Let f be a circulation and g a potential difference in D, and let T be a spanning tree of D. Show that f is uniquely determined by $f\mid\bar{T}$ and g by $g\mid T$.

12.1.2 (*a*) Let **B** and **C** be basis matrices of $\mathscr{B}$ and $\mathscr{C}$ and let T be any spanning tree of D. Show that **B** is uniquely determined by $\mathbf{B}\mid T$ and **C** is uniquely determined by $\mathbf{C}\mid\bar{T}$.

(*b*) Let T and T_1 be two fixed spanning trees of D. Let $\mathbf{B}$ and $\mathbf{B}_1$ denote the basis matrices of $\mathscr{B}$, and $\mathbf{C}$ and $\mathbf{C}_1$ the basis matrices of $\mathscr{C}$, corresponding to the trees T and T_1. Show that $\mathbf{B} = (\mathbf{B} \mid T_1)\mathbf{B}_1$ and $\mathbf{C} = (\mathbf{C} \mid \bar{T}_1)\mathbf{C}_1$.

12.1.3 Let $\mathbf{K}$ denote the matrix obtained from the incidence matrix $\mathbf{M}$ of a connected digraph D by deleting any one of its rows. Show that $\mathbf{K}$ is a basis matrix of $\mathscr{B}$.

12.1.4 Show that if G is a plane graph, then $\mathscr{B}(G) \cong \mathscr{C}(G^*)$ and $\mathscr{C}(G) \cong \mathscr{B}(G^*)$.

12.1.5 A *circulation* of D *over a field* F is a function $f: A \to F$ which satisfies (12.1) in F; a *potential difference* of D *over* F is similarly defined. The vector spaces of these potential differences and circulations are denoted by $\mathscr{B}_F$ and $\mathscr{C}_F$. Show that theorem 12.2 remains valid if $\mathscr{B}$ and $\mathscr{C}$ are replaced by $\mathscr{B}_F$ and $\mathscr{C}_F$, respectively.

12.2 THE NUMBER OF SPANNING TREES

In this section we shall derive a formula for the number of spanning trees in a graph.

Let G be a connected graph and let T be a fixed spanning tree of G. Consider an arbitrary orientation D of G and let $\mathbf{B}$ be the basis matrix of $\mathscr{B}$ corresponding to T. It follows from theorem 12.2 that if S is a subset of A with $|S| = \nu - 1$ then the square submatrix $\mathbf{B} \mid S$ is nonsingular if and only if S is a spanning tree of G. Thus the number of spanning trees of G is equal to the number of nonsingular submatrices of $\mathbf{B}$ of order $\nu - 1$.

A matrix is said to be *unimodular* if all its full square submatrices have determinants 0, +1 or −1. The proof of the following theorem is due to Tutte (1965*b*).

Theorem 12.3 The basis matrix $\mathbf{B}$ is unimodular.

Proof Let $\mathbf{P}$ be a full submatrix of $\mathbf{B}$ (one of order $\nu - 1$). Suppose that $\mathbf{P} = \mathbf{B} \mid T_1$. We may assume that T_1 is a spanning tree of D since, otherwise, $\det \mathbf{P} = 0$ by theorem 12.2. Let $\mathbf{B}_1$ denote the basis matrix of $\mathscr{B}$ corresponding to T_1. Then (exercise 12.1.2*b*)

$$(\mathbf{B} \mid T_1)\mathbf{B}_1 = \mathbf{B}$$

Restricting both sides to T, we obtain

$$(\mathbf{B} \mid T_1)(\mathbf{B}_1 \mid T) = \mathbf{B} \mid T$$

Noting that $\mathbf{B} \mid T$ is an identity matrix, and taking determinants, we get

$$\det(\mathbf{B} \mid T_1)\det(\mathbf{B}_1 \mid T) = 1 \tag{12.5}$$

Both determinants in (12.5), being determinants of integer matrices, are themselves integers. It follows that $\det(\mathbf{B} \mid T_1) = \pm 1$ □

Theorem 12.4 $\tau(G) = \det \mathbf{BB}'$ (12.6)

Proof Using the formula for the determinant of the product of two rectangular matrices (see Hadley, 1961), we obtain

$$\det \mathbf{BB}' = \sum_{\substack{S \subseteq A \\ |S| = \nu - 1}} (\det(\mathbf{B} \mid S))^2 \tag{12.7}$$

Now, by theorem 12.2, the number of nonzero terms in (12.7) is equal to $\tau(G)$. But, by theorem 12.3, each such term has value 1 □

One can similarly show that if $\mathbf{C}$ is a basis matrix of $\mathscr{C}$ corresponding to a tree, then $\mathbf{C}$ is unimodular and

$$\tau(G) = \det \mathbf{CC}' \tag{12.8}$$

Corollary 12.4 $\tau(G) = \pm \det \begin{bmatrix} \mathbf{B} \\ \hline \mathbf{C} \end{bmatrix}$

Proof By (12.6) and (12.8)

$$(\tau(G))^2 = \det \mathbf{BB}' \det \mathbf{CC}' = \det \left[\begin{array}{c|c} \mathbf{BB}' & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{CC}' \end{array}\right]$$

Since $\mathscr{B}$ and $\mathscr{C}$ are orthogonal, $\mathbf{BC}' = \mathbf{CB}' = \mathbf{0}$. Thus

$$(\tau(G))^2 = \det \left[\begin{array}{c|c} \mathbf{BB}' & \mathbf{BC}' \\ \hline \mathbf{CB}' & \mathbf{CC}' \end{array}\right] = \det \left(\begin{bmatrix} \mathbf{B} \\ \hline \mathbf{C} \end{bmatrix} [\mathbf{B}' \mid \mathbf{C}'] \right)$$

$$= \det \begin{bmatrix} \mathbf{B} \\ \hline \mathbf{C} \end{bmatrix} \det [\mathbf{B}' \mid \mathbf{C}'] = \left(\det \begin{bmatrix} \mathbf{B} \\ \hline \mathbf{C} \end{bmatrix} \right)^2$$

The corollary follows on taking square roots □

Since theorem 12.2 is valid for all basis matrices of $\mathscr{B}$, (12.6) clearly holds for any such matrix $\mathbf{B}$ that is unimodular. In particular, a matrix $\mathbf{K}$ obtained by deleting any one row of the incidence matrix $\boldsymbol{M}$ is unimodular (exercise 12.2.1*a*). Thus

$$\tau(G) = \det \mathbf{KK}'$$

This expression for the number of spanning trees in a graph is implicit in the work of Kirchhoff (1847), and is known as the *matrix-tree theorem.*

Exercises

12.2.1 Show that

(*a*)* a matrix **K** obtained from ***M*** by deleting any one row is unimodular;

(*b*) $\tau(G) = \pm\det\begin{bmatrix}\mathbf{K}\\---\\\mathbf{C}\end{bmatrix}$

12.2.2 The *conductance matrix* $\mathbf{C} = [c_{ij}]$ of a loopless graph G is the $\nu \times \nu$ matrix in which

$$c_{ii} = \sum_{j \neq i} a_{ij} \quad \text{for all} \quad i$$

$$c_{ij} = -a_{ij} \quad \text{for all} \quad i \quad \text{and} \quad j \quad \text{with} \quad i \neq j$$

where $\mathbf{A} = [a_{ij}]$ is the adjacency matrix of G. Show that

(*a*) $\mathbf{C} = \mathbf{M}\mathbf{M}'$, where ***M*** is the incidence matrix of any orientation of G;

(*b*) all cofactors of **C** are equal to $\tau(G)$.

12.2.3 A matrix is *totally unimodular* if all square submatrices have determinants 0, +1 or −1. Show that

(*a*) any basis matrix of $\mathscr{B}$ or $\mathscr{C}$ corresponding to a tree is totally unimodular;

(*b*) the incidence matrix of a simple graph G is totally unimodular if and only if G is bipartite.

12.2.4 Let F be a field of characteristic p. Show that

(*a*) if **B** and **C** are basis matrices of $\mathscr{B}_F$ and $\mathscr{C}_F$, respectively, corresponding to a tree, then $\det\begin{bmatrix}\mathbf{B}\\---\\\mathbf{C}\end{bmatrix} = \pm\,\tau(G) \pmod p$;

(*b*) $\dim(\mathscr{B}_F \cap \mathscr{C}_F) > 0$ if and only if $p \mid \tau(G)$. (H. Shank)

APPLICATIONS

12.3 PERFECT SQUARES

A *squared rectangle* is a rectangle dissected into at least two (but a finite number of) squares. If no two of the squares in the dissection have the same size, then the squared rectangle is *perfect*. The *order* of a squared rectangle is the number of squares into which it is dissected. Figure 12.7 shows a perfect rectangle of order 9. A squared rectangle is *simple* if it does not contain a rectangle which is itself squared. Clearly, every squared rectangle is composed of ones that are simple.

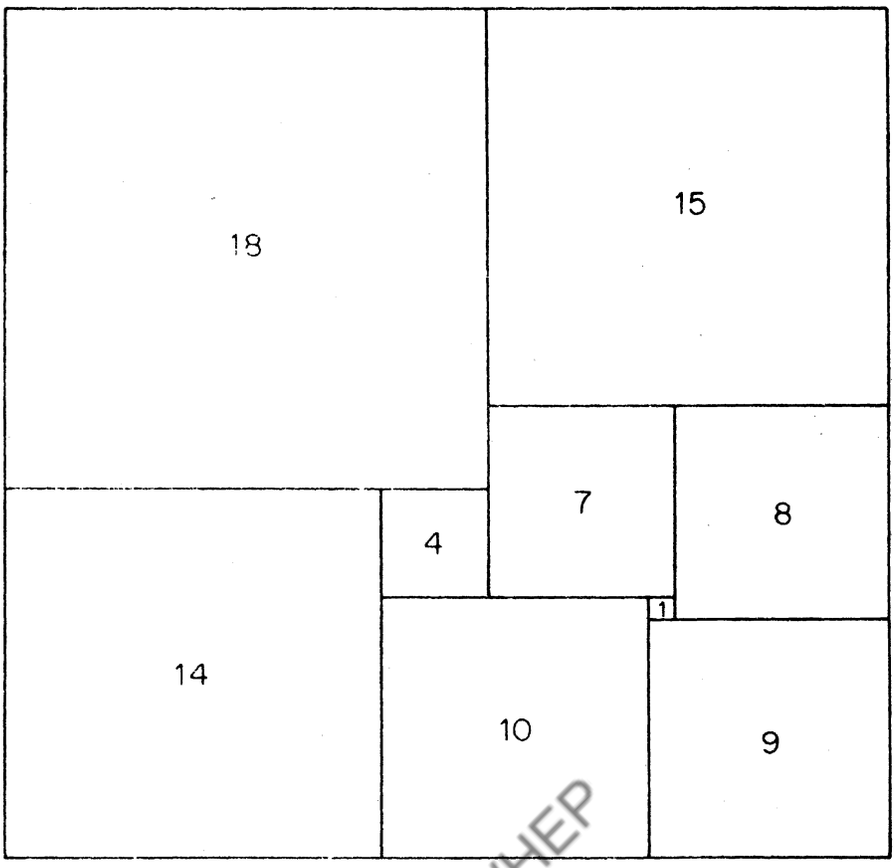
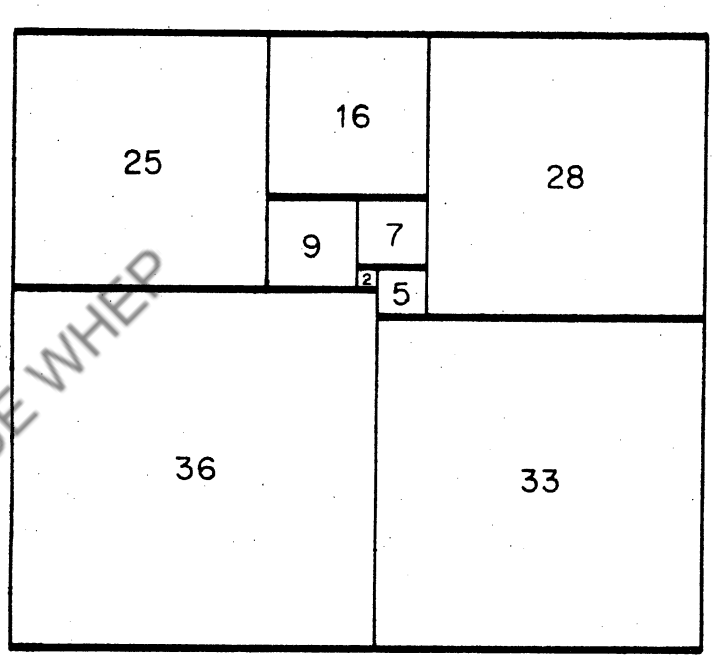


Figure 12.7. A perfect rectangle

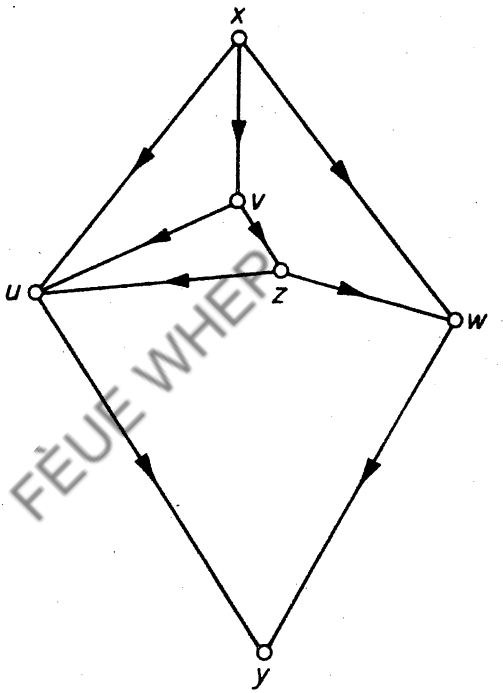
For a long time no perfect squares were known, and it was conjectured that such squares did not exist. Sprague (1939) was the first to publish an example of a perfect square. About the same time, Brooks et al. (1940) developed systematic methods for their construction by using the theory of graphs. In this section, we shall present a brief discussion of their methods.

We first show how a digraph can be associated with a given squared rectangle R. The union of the horizontal sides of the constituent squares in the dissection consists of horizontal line segments; each such segment is called a *horizontal dissector* of R. In figure 12.8*a*, the horizontal dissectors are indicated by solid lines. We can now define the digraph D associated with R. To each horizontal dissector of R there corresponds a vertex of D; two vertices v_i and v_j of D are joined by an arc (v_i, v_j) if and only if their corresponding horizontal dissectors H_i and H_j flank some square of the dissection and H_i lies above H_j in R. Figure 12.8*b* shows the digraph associated with the squared rectangle in figure 12.8*a*. The vertices corresponding to the upper and lower sides of R are called the *poles* of D and are denoted by x and y, respectively.

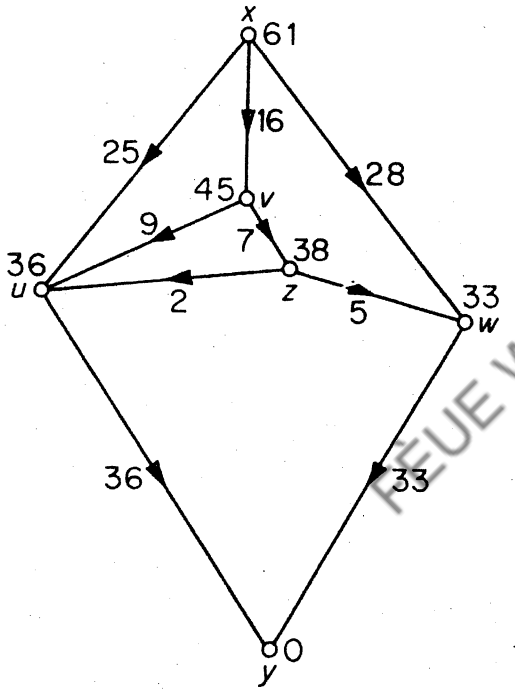
We now assign to each vertex v of D a potential $p(v)$ equal to the height (above the lower side of R) of the corresponding horizontal dissector. If we regard D as an electrical network in which each wire has unit resistance, the potential difference $g = \delta p$ determines a flow of currents from x to y (see



(a)



(b)



(c)

Figure 12.8

figure 12.8c). These currents satisfy *Kirchhoff's current law:* the total amount of current entering a vertex $v \in V\backslash\{x, y\}$ is equal to the total amount leaving it. For example, the total amount entering u in figure 12.8c is $25+9+2=36$, and the same amount leaves this vertex.

Let D be the digraph corresponding to a squared rectangle R, with poles x and y, and let G be the underlying graph of D. Then the graph $G+xy$ is called the *horizontal graph* of R. Brooks et al. (1940) showed that the horizontal graph of any simple squared rectangle is a 3-connected planar graph (their definition of connectivity differs slightly from the one used in this book). They also showed that, conversely, if H is a 3-connected planar graph and $xy \in E(H)$, then any flow of currents from x to y in $H-xy$ determines a squared rectangle. Thus one possible way of searching for perfect rectangles of order n is to

(i) list all 3-connected planar graphs with $n+1$ edges, and
(ii) for each such graph H and each edge xy of H, determine a flow of currents from x to y in $H-xy$.

Tutte (1961) showed that every 3-connected planar graph can be derived from a wheel by a sequence of operations involving face subdivisions and the taking of duals. Bouwkamp, Duijvestijn and Medema (1960) then applied Tutte's theorem to list all 3-connected planar graphs with at most 16 edges. Here we shall see how the theory developed in sections 12.1 and 12.2 can be used in computing a flow of currents from x to y in a digraph D.

Let $g(a)$ denote the current in arc a of D, and suppose that the total current leaving x is σ. Then

$$\sum_{a\in A} m_x(a)g(a) = \sigma \tag{12.9}$$

Kirchhoff's current law can be formulated as

$$\sum_{a\in A} m_v(a)g(a) = 0 \quad \text{for all} \quad v \in V\backslash\{x, y\} \tag{12.10}$$

Now, since g is a potential difference, it is orthogonal to every circulation. Therefore,

$$\mathbf{C}g' = \mathbf{0} \tag{12.11}$$

where $\mathbf{C}$ is a basis matrix of $\mathscr{C}$ corresponding to a tree T of D and g' is the transpose of the vector g. Equations (12.9)–(12.11) together give the matrix equation

$$\begin{bmatrix}\mathbf{K}\\ \cdots \\ \mathbf{C}\end{bmatrix} g' = \begin{bmatrix}\sigma\\ \cdots \\ \mathbf{0}\end{bmatrix} \tag{12.12}$$

where $\mathbf{K}$ is the matrix obtained from $\boldsymbol{M}$ by deleting the row m_y. This

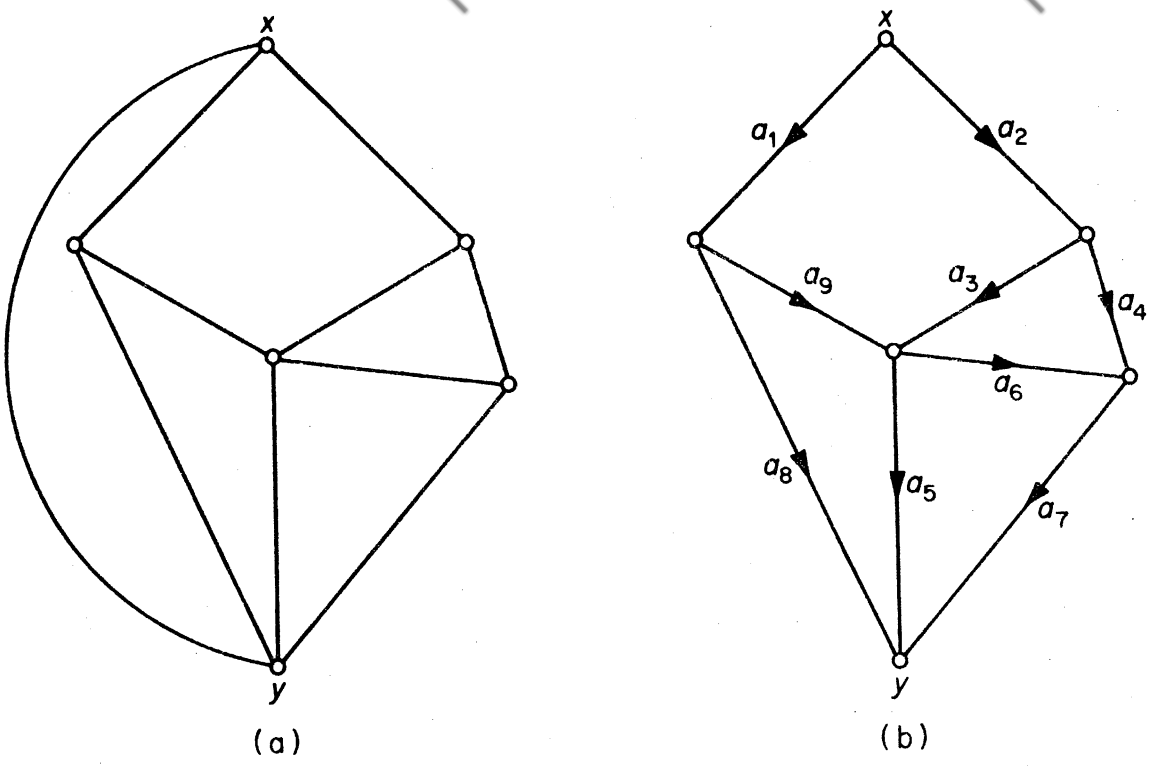


Figure 12.9

equation can be solved using Cramér's rule. Note that, since $\det \begin{bmatrix} \mathbf{K} \\ \hline \mathbf{C} \end{bmatrix} = \pm\tau(G)$ (exercise 12.2.1*b*), we obtain a solution in integers if $\sigma = \tau(G)$. Thus, in computing the currents, it is convenient to take the total current leaving x to be equal to the number of spanning trees of D.

We illustrate the above procedure with an example. Consider the 3-connected planar graph in figure 12.9*a*. On deleting the edge xy and orienting each edge we obtain the digraph D of figure 12.9*b*.

It can be checked that the number of spanning trees in D is 66. By considering the tree $T = \{a_1, a_2, a_3, a_4, a_5\}$ we obtain the following nine equations, as in (12.12), (with $g(a_i)$ written simply as g_i).

$$
\begin{aligned}
g_1 + g_2 &= 66 \\
g_1 - g_8 - g_9 &= 0 \\
g_2 - g_3 - g_4 &= 0 \\
g_3 - g_5 - g_6 + g_9 &= 0 \\
g_4 + g_6 - g_7 &= 0 \\
g_3 - g_4 + g_6 &= 0 \\
-g_3 + g_4 - g_5 + g_7 &= 0 \\
g_1 - g_2 - g_3 - g_5 + g_8 &= 0 \\
g_1 - g_2 - g_3 + g_9 &= 0
\end{aligned}
$$

The solution to this system of equations is given by

$$(g_1, g_2, g_3, g_4, g_5, g_6, g_7, g_8, g_9) = (36, 30, 14, 16, 20, 2, 18, 28, 8)$$

The squared rectangle based on this flow of currents is just the one in figure 12.7 with all dimensions doubled.

Figure 12.10 shows a simple perfect square of order 25. It was discovered by Wilson (1967), and is the smallest (least order) such square known.

Further results on perfect squares can be found in the survey article by Tutte (1965*a*).

Exercises

12.3.1 Show that the constituent squares in a squared rectangle have commensurable sides.

12.3.2 The *vertical graph* of a squared rectangle R is the horizontal graph of the squared rectangle obtained by rotating R through a right angle. If no point of R is the corner of four constituent squares, show that the horizontal and vertical graphs of R are duals.

12.3.3* A *perfect cube* is a cube dissected into a finite number of smaller cubes, no two of the same size. Show that there exists no perfect cube.

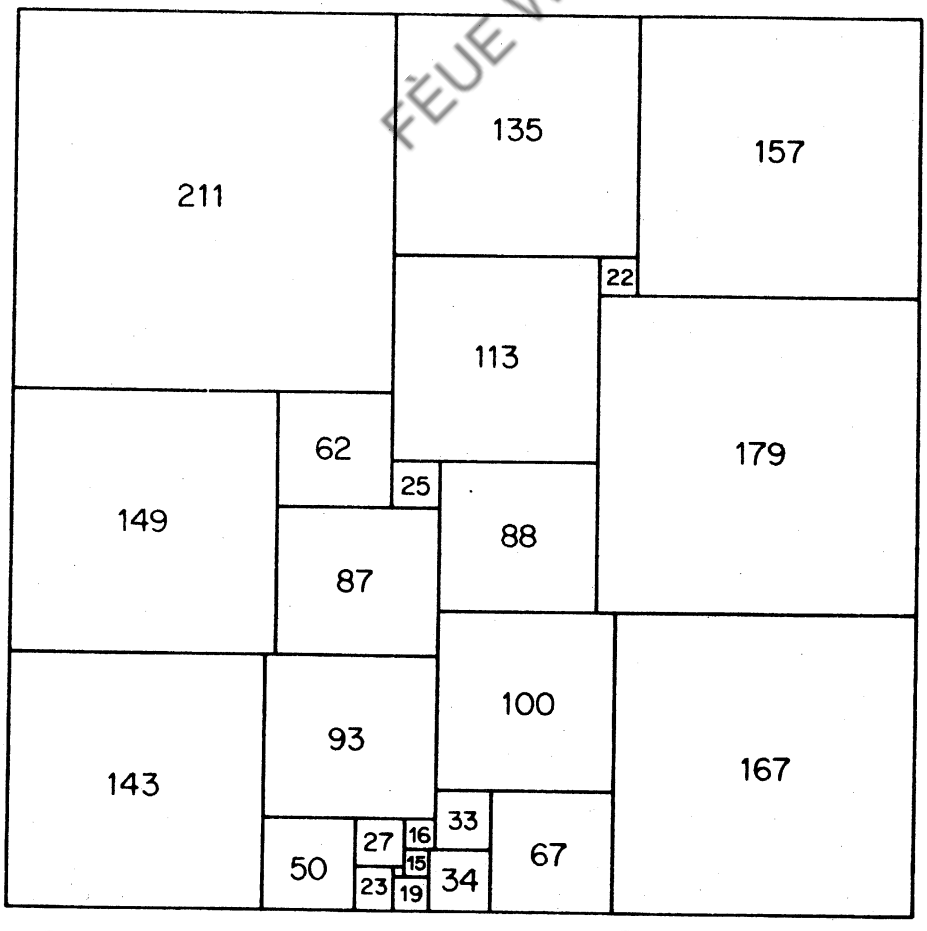


Figure 12.10. A simple perfect square of order 25

REFERENCES

Bouwkamp, C. J., Duijvestijn, A. J. W. and Medema, P. (1960). *Tables Relating to Simple Squared Rectangles of Orders Nine through Fifteen*, Technische Hogeschool, Eindhoven

Brooks, R. L., Smith, C. A. B., Stone, A. H. and Tutte, W. T. (1940). The dissection of rectangles into squares. *Duke Math. J.*, **7,** 312–40

Hadley, G. (1961). *Linear Algebra*, Addison-Wesley, Reading, Mass.

Kirchhoff, G. (1847). Über die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Verteilung galvanischer Ströme geführt wird, *Ann. Phys. Chem.*, **72,** 497–508

Sprague, R. (1939). Beispiel einer Zerlegung des Quadrats in lauter verschiedene Quadrate, *Math. Z.*, **45,** 607–8

Tutte, W. T. (1961). A theory of 3-connected graphs. *Nederl. Akad. Wetensch. Proc. Ser. A.*, **23,** 441–55

Tutte, W. T. (1965*a*). The quest of the perfect square, *Amer. Math. Monthly*, **72,** 29–35

Tutte, W. T. (1965*b*). Lectures on matroids. *J. Res. Nat. Bur. Standards Sect. B*, **69,** 1–47

Tutte, W. T. (1971). *Introduction to Matroid Theory*, Elsevier, New York

Wilson, J. C. (1967). *A Method for Finding Simple Perfect Square Squarings.* Ph.D. Thesis, University of Waterloo

Appendix I
Hints to Starred Exercises

1.2.9(*b*) If $G \not\cong T_{m,n}$, then G has parts of size $n_1, n_2, \ldots, n_m$, with $n_i - n_j > 1$ for some i and j. Show that the complete m-partite graph with parts of size $n_1, n_2, \ldots, n_i - 1, \ldots, n_j + 1, \ldots, n_m$ has more edges than G.

1.3.3 In terms of the adjacency matrix $\mathbf{A}$, an automorphism of G is a permutation matrix $\mathbf{P}$ such that $\mathbf{PAP}' = \mathbf{A}$ or, equivalently, $\mathbf{PA} = \mathbf{AP}$ (since $\mathbf{P}' = \mathbf{P}^{-1}$). Show that if $\mathbf{x}$ is an eigenvector of $\mathbf{A}$ belonging to an eigenvalue λ, then, for any automorphism $\mathbf{P}$ of G, so is $\mathbf{Px}$. Since the eigenvalues of $\mathbf{A}$ are distinct and $\mathbf{P}$ is orthogonal, $\mathbf{P}^2\mathbf{x} = \mathbf{x}$ for all eigenvectors $\mathbf{x}$.

1.4.5 Suppose that all induced subgraphs of G on n vertices have m edges. Show that, for any two vertices v_i and v_j,

$$\varepsilon(G) - d(v_i) = \varepsilon(G - v_i) = m\binom{\nu - 1}{n} \Big/ \binom{\nu - 3}{n - 2}$$

$$\varepsilon(G) - d(v_i) - d(v_j) + a_{ij} = \varepsilon(G - v_i - v_j) = m\binom{\nu - 2}{n} \Big/ \binom{\nu - 4}{n - 2}$$

where $a_{ij} = 1$ or 0 according as v_i and v_j are adjacent or not. Deduce that a_{ij} is independent of i and j.

1.5.7(*a*) To prove the necessity, first show that if G is simple with u_1v_1, $u_2v_2 \in E$ and u_1v_2, $u_2v_1 \notin E$, then $G - \{u_1v_1, u_2v_2\} + \{u_1v_2, u_2v_1\}$ has the same degree sequence as G. Using this, show that if $\mathbf{d}$ is graphic, then there is a simple graph G with $V = \{v_1, v_2, \ldots, v_n\}$ such that (i) $d(v_i) = d_i$ for $1 \le i \le n$, and (ii) v_1 is joined to $v_2, v_3, \ldots, v_{d_1+1}$. The graph $G - v_1$ has degree sequence $\mathbf{d}'$.

1.5.8 Show that a bipartite subgraph with the largest possible number of edges has this property.

1.5.9 Define a graph on S in which x_i and x_j are adjacent if and only if they are at distance one. Show that in this graph each vertex has degree at most six.

1.7.3 Consider a longest path and the vertices adjacent to the origin of this path.

1.7.6(*b*) By contradiction. Let G be a smallest counter-example. Show that (i) the girth of G is at least five, and (ii) $\delta \ge 3$. Deduce that $\nu \le 8$ and show that no such graph exists.

2.1.10 To prove the sufficiency, consider a graph G with degree sequence $\mathbf{d} = (d_1, d_2, \ldots, d_\nu)$ and as few components as possible. If

G is not connected, show that, by a suitable exchange of edges (as in the hint to exercise 1.5.7*a*), there is a graph with degree sequence **d** and fewer components than G.

2.2.12 Define a labelled graph G as follows: the vertices of G are the subsets $A_1, A_2, \ldots, A_n$, and A_i is joined to A_j $(i \neq j)$ by an edge labelled a if either $A_i = A_j \cup \{a\}$ or $A_j = A_i \cup \{a\}$. For any subgraph H of G, let $L(H)$ be the set of labels on edges of H. Show that if F is a maximal forest of G, then $L(F) = L(G)$. Any element x in $S \backslash L(F)$ has the required property.

2.4.2 Several applications of theorem 2.8 yield the recurrence relation

$$w_n - 4w_{n-1} + 4w_{n-2} - 1 = 0$$

where w_n is the number of spanning trees in the wheel with n spokes. Solve this recurrence relation.

3.2.6 Form a new graph G' by adding two vertices x and y, and joining x to all vertices in X and y to all vertices in Y. Show that G' is 2-connected and apply theorem 3.2.

3.2.7(*a*) Use induction on ε. Let $e_1 \in E$. If $G \cdot e_1$ is a critical block, then $G \cdot e_1$ has a vertex of degree two and, hence, so does G. If $G \cdot e_1$ is not critical, there is an $e_2 \in E \backslash \{e_1\}$ such that $(G \cdot e_1) - e_2$ is a block. Using the fact that $(G \cdot e_1) - e_2 = (G - e_2) \cdot e_1$, show that e_1 and e_2 are incident with a vertex of degree two in G.

(*b*) Use (*a*) and induction on ν.

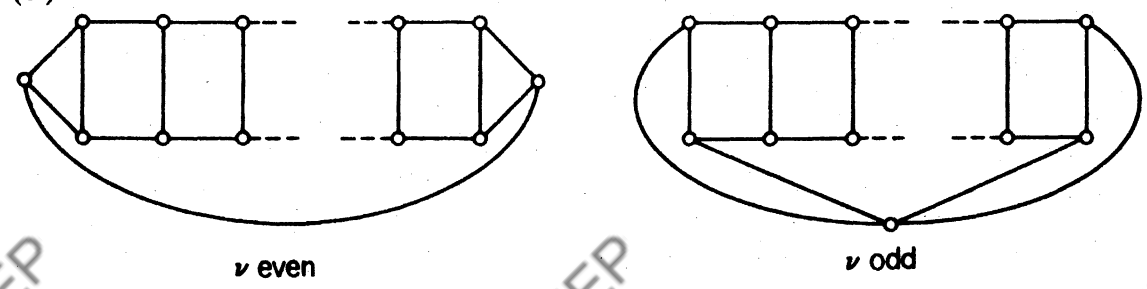
4.1.6 Necessity: if $G - v$ contains a cycle C, consider an Euler tour (with origin v) of the component of $G - E(C)$ that contains v. Sufficiency: let Q be a (v, w)-trail of G which is not an Euler tour. Show that $G - E(Q)$ has exactly one nontrivial component.

4.2.4 Form a new graph G' by adding a new vertex and joining it to every vertex of G. Show that G has a Hamilton path if and only if G' has a Hamilton cycle, and apply theorem 4.5.

4.2.6 Form a new graph G' by adding edges so that $G'[X]$ is complete. Show that G is hamiltonian if and only if G' is hamiltonian, and apply theorem 4.5.

4.2.9 Let P be a longest path in G. If P has length $l < 2\delta$, show, using the proof technique of theorem 4.3, that G has a cycle of length $l + 1$. Now use the fact that G is connected to obtain a contradiction.

4.2.11(*b*)



4.2.13 Use the fact that the Petersen graph is hypohamiltonian (exercise 4.2.12).

4.4.1 Consider an Euler tour Q in the weighted graph formed from T by duplicating each of its edges. Now make use of triangle inequalities to obtain from Q a Hamilton cycle in G of weight at most $w(Q)$.

5.1.5(a) To show that K_{2n} is 1-factorable, arrange the vertices in the form of a regular $(2n-1)$-gon with one vertex in the centre. A radial edge together with the edges perpendicular to it is a perfect matching.

5.1.6 Label the vertices $0, 1, 2, \ldots, 2n$ and arrange the vertices $1, 2, \ldots, 2n$ in a circle with 0 at the centre. Let $C = (0, 1, 2, 2n, 3, 2n-1, 4, 2n-2, \ldots, n+2, n+1, 0)$ and consider the rotations of C.

5.2.3(b) Let G be a $2k$-regular graph with $V = \{v_1, v_2, \ldots, v_\nu\}$; without loss of generality, assume that G is connected. Let C be an Euler tour in G. Form a bipartite graph G' with bipartition (X, Y), where $X = \{x_1, x_2, \ldots, x_\nu\}$ and $Y = \{y_1, y_2, \ldots, y_\nu\}$ by joining x_i to y_j whenever v_i immediately precedes v_j on C. Show that G' is 1-factorable and hence that G is 2-factorable.

5.2.8 Construct a bipartite graph G with bipartition (X, Y) in which X is the set of rows of $\mathbf{Q}$, Y is the set of columns of $\mathbf{Q}$, and row i is joined to column j if and only if the entry q_{ij} is positive. Show that G has a perfect matching, and then use induction on the number of nonzero entries of $\mathbf{Q}$.

5.3.1 Let G be a bipartite graph with bipartition (X, Y). Assume that ν is even (the case when ν is odd requires a little modification). Obtain a graph H from G by joining all pairs of vertices in Y. G has a matching that saturates every vertex in X if and only if H has a perfect matching.

5.3.4 Let G^* be a maximal spanning supergraph of G such that the number of edges in a maximum matching of G^* is the same as for G. Show, using the proof technique ot theorem 5.4, that if U is the set of vertices of degree $\nu-1$ in G^* then G^*-U is a disjoint union of complete graphs.

6.2.1 See the hint to exercise 5.1.5a.

6.2.8 Use the proof technique of theorem 6.2.

7.1.3(b) Let $v_1v_2 \ldots v_n$ be a longest path in G. Show that $G-v_2$ has at most one nontrivial component, and use induction on ε.

7.2.6(b) Let $p(m-1) = n-1$. The complete $(p+1)$-partite graph with $m-1$ vertices in each part shows that $r(T, K_{1,n}) > (p+1)(m-1) = m+n-2$. To prove that $r(T, K_{1,n}) \leq m+n-1$, show that any simple graph G with $\delta \geq m-1$ contains every tree T on m vertices.

(*c*) The complete $(n-1)$-partite graph with $m-1$ vertices in each part shows that $r(T, K_n) > (m-1)(n-1)$. To prove that $r(T, K_n) \leq (m-1)(n-1)+1$, use induction on n and the fact that any simple graph with $\delta \geq m-1$ contains every tree T on m vertices.

7.3.3(*c*) Assume G contains no triangle. Choose a shortest odd cycle C in G. Show that each vertex in $V(G) \backslash V(C)$ can be joined to at most two vertices of C. Apply exercise 7.3.3*a* to $G - V(C)$, and obtain a contradiction.

7.3.4(*a*) G contains $K_{2,m}$ if and only if there are m vertices with a pair of common neighbours. Any vertex v has $\binom{d(v)}{2}$ pairs of neighbours. Therefore if $\sum_{v \in V} \binom{d(v)}{2} > (m-1)\binom{v}{2}$, G contains $K_{2,m}$.

7.5.1 Define a graph G by $V(G) = \{x_1, \ldots, x_n\}$, and $E(G) = \{x_i x_j \mid d(x_i, x_j) = 1\}$, and show that if all edges of G are drawn as straight line segments, then (i) any two edges of G are either adjacent or cross, and (ii) if some vertex of G has degree greater than two, it is adjacent to a vertex of degree one. Then prove (a) by induction on n.

8.1.o Let $\mathscr{C} = (V_1, V_2, \ldots, V_k)$ be a k-colouring of G, and let $\mathscr{C}'$ be a colouring of G in which each colour class contains at least two vertices. If $|V_i| \geq 2$ for all i, there is nothing to prove, so assume that $V_1 = \{v_1\}$. Let $u_2 \in V_2$ be a vertex of the same colour as v_1 in $\mathscr{C}'$. Clearly $|V_2| \geq 2$. If $|V_2| > 2$, transfer u_2 to V_1. Otherwise, let v_2 be the other vertex in V_2. In $\mathscr{C}'$, v_1 and v_2 must be assigned different colours. Let $u_3 \in V_3$ be a vertex of the same colour as v_2 in $\mathscr{C}'$. As before, $|V_3| \geq 2$. Proceeding in this way, one must eventually find a set V_i with $|V_i| > 2$. G can now be recoloured so that fewer colour classes contain only one vertex.

8.1.13(*a*) Let $(X_1, X_2, \ldots, X_n)$ and $(Y_1, Y_2, \ldots, Y_n)$ be n-colourings of $G[X]$ and $G[Y]$, respectively. Construct a bipartite graph H with bipartition $(\{x_1, x_2, \ldots, x_n\}, \{y_1, y_2, \ldots, y_n\})$ by joining x_i and y_j if and only if the edge cut $[X_i, Y_j]$ is empty in G. Using exercise 5.2.6*b*, show that H has a perfect matching. If x_i is matched with $y_{f(i)}$ under this matching, let $V_i = X_i \cup Y_{f(i)}$. Show that $(V_1, V_2, \ldots, V_n)$ is an n-colouring of G.

8.3.1 Show that it suffices to consider 2-connected graphs. Choose a longest cycle C in G and show that there are two paths across C as in theorem 8.5.

8.3.2(*a*) If $\delta \geq 3$, use exercise 8.3.1. If there is a vertex of degree less than three, delete it and use induction.

8.4.8 Consider the expansion of $\pi_k(G)$ in terms of chromatic polynomials of complete graphs.

8.5.2(*a*) It is easily verified that H has girth at least six. If H is k-colourable, there is a ν-element subset of S all of whose members receive the same colour. Consider the corresponding copy of G and obtain a contradiction.

9.2.8 The dual G^* is 2-edge-connected and 3-regular and, hence (corollary 5.4), has a perfect matching M. $(G^* \cdot M)^*$ is a bipartite subgraph of G.

10.2.2 Form a new digraph on the same vertex set joining u to v if v is reachable from u, and apply corollary 10.1.

10.2.5 Let D_1 and D_2 be the spanning subdigraphs of D such that the arcs of D_1 are the arcs (u, v) of D for which $f(u) \leq f(v)$, and the arcs of D_2 are the arcs (u, v) for which $f(u) > f(v)$. Show that either $\chi(D_1) > m$ or $\chi(D_2) > n$, and apply theorem 10.1.

10.3.4 Let $v_1 v_2 \ldots v_{2n+1} v_1$ be an odd cycle. If $(v_i, v_{i+1}) \in A$, set $P_i = (v_i, v_{i+1})$; if $(v_i, v_{i+1}) \notin A$, let P_i be a directed (v_i, v_{i+1})-path. If some P_i is of even length, $P_i + (v_{i+1}, v_i)$ is a directed odd cycle; otherwise, $P_1 P_2 \ldots P_{2n+1}$ is a closed directed trail of odd length, and therefore contains a directed odd cycle.

11.3.5 Use the construction given in the proof of theorem 11.6, and assign capacity $m(v)$ to arc (v', v'').

11.4.4 Use induction on k and exercise 11.4.3.

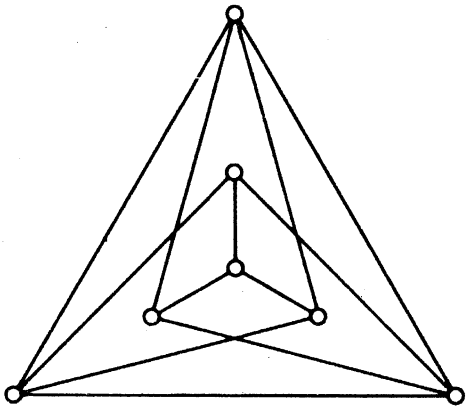
11.5.4 Use an argument similar to that in exercise 1.5.7.

11.5.5(*a*) Necessity follows on taking V_1 as the set of vertices with indegree m and V_2 as the set of vertices with indegree n. To prove sufficiency, construct a network N by forming the associated digraph of G, assigning unit capacity to each arc, and regarding the elements of V_1 as sources and the elements of V_2 as sinks. By theorem 11.8, there is a flow f in N (which can be assumed integral) in which the supply at each source and demand at each sink is $|m-n|$. The f-saturated arcs induce an (m,n)-orientation on a subgraph H of G. An (m,n)-orientation of G can now be obtained by giving the remaining edges an eulerian orientation.

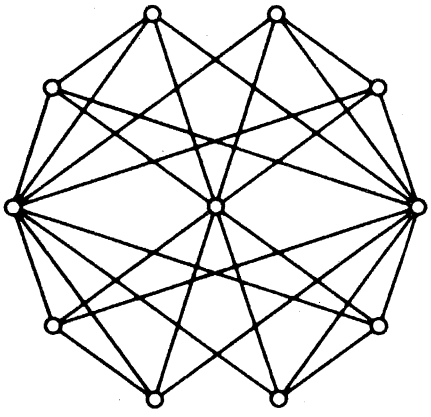
12.2.1(*a*) Use induction on the order of the submatrix. Let $\mathbf{P}$ be a square submatrix. If each column of $\mathbf{P}$ contains two nonzero entries, then $\det \mathbf{P} = 0$. Otherwise, expand $\det \mathbf{P}$ about a column with exactly one nonzero entry, and apply the induction hypothesis.

12.3.3 Show, first, that in any perfect rectangle the smallest constituent square is not on the boundary of the rectangle. Now suppose that there is a perfect cube and consider the perfect square induced on the base of this cube by the constituent cubes.

Appendix II
Four Graphs and a Table of their Properties

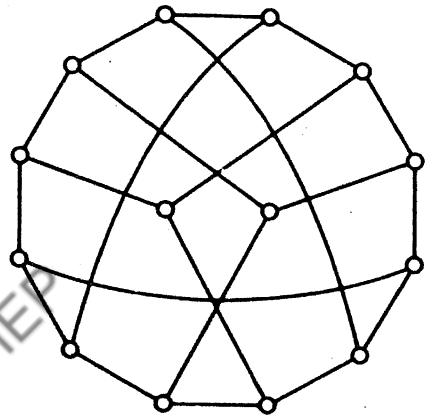


G_1

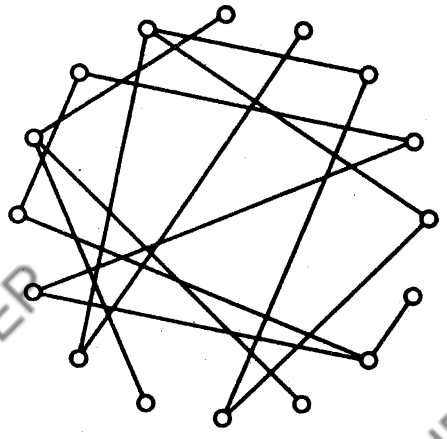


G_2

	ν	ε	δ	Δ	ω	κ	κ	α	α'	β	β'	χ	χ'
G_1	7	12	3	4	1	3	3	3	3	4	4	4	4
G_2	11	28	4	8	1	3	4	4	5	7	6	3	8
G_3	14	21	3	3	1	3	3	7	7	7	7	2	3
G_4	16	15	1	3	3	0	0	9	7	7	9	3	3



G_3



G_4

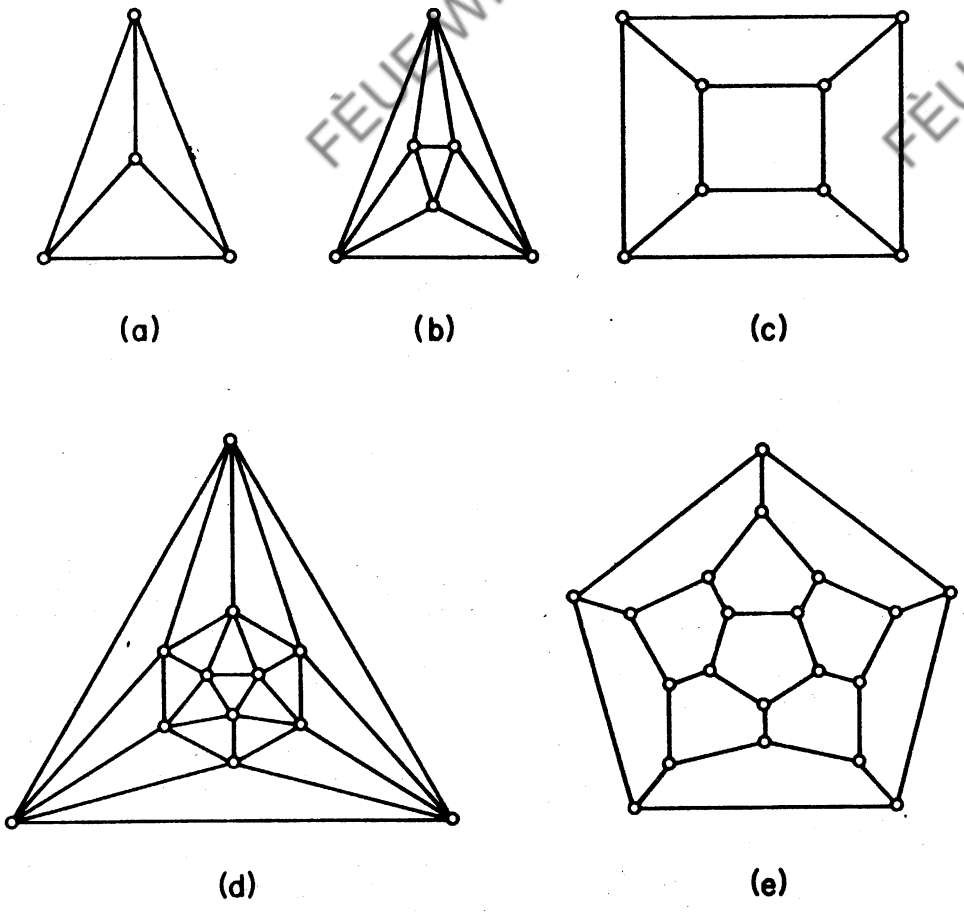
diameter	girth	bipartite?	eulerian?	hamiltonian?	critical?	planar?
2	3	No	No	Yes	Yes	Yes
2	3	No	Yes	No	No	No
3	6	Yes	No	Yes	No	No
∞	4	No	No	No	No	Yes

Appendix III
Some Interesting Graphs

There are a number of graphs which are interesting because of their special structure. We have already met some of these (for example, the Grinberg graph, the Grötzsch graph, the Herschel graph and the Ramsey graphs). Here we present a selection of other interesting graphs and families of graphs.

THE PLATONIC GRAPHS

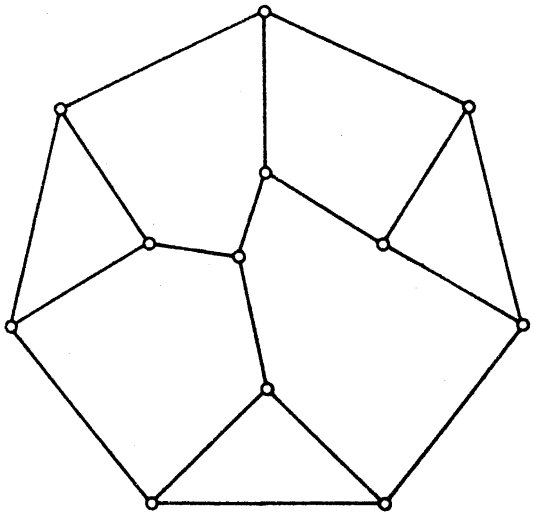
These are the graphs whose vertices and edges are the vertices and edges of the platonic solids (see Fréchet and Fan, 1967).



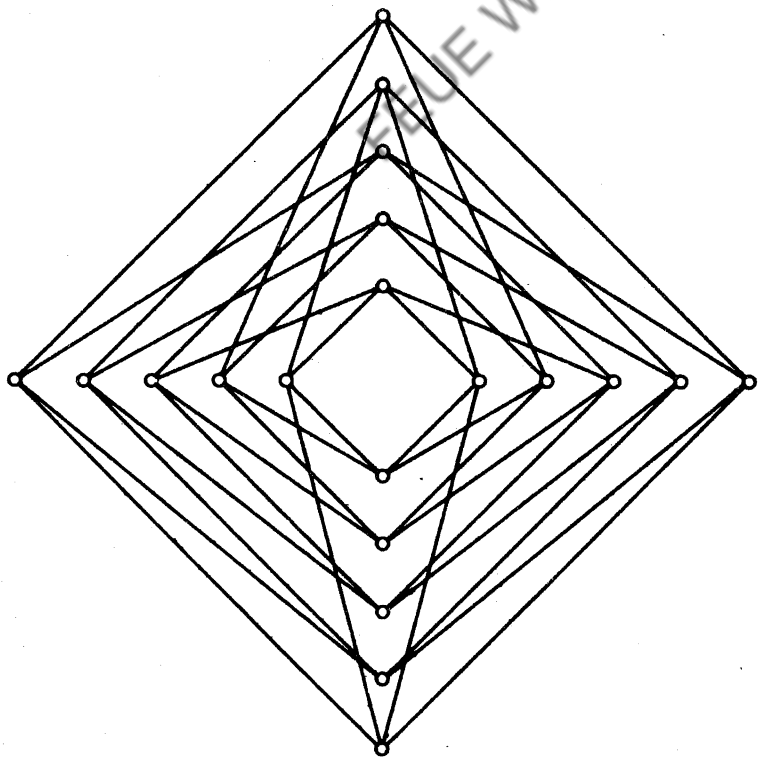
(*a*) The tetrahedron; (*b*) the octahedron; (*c*) the cube; (*d*) the icosahedron; (*e*) the dodecahedron

AUTOMORPHISM GROUPS

(i) As has already been noted (exercise 1.2.12), every group is isomorphic to the automorphism group of some graph. Frucht (1949) showed, in fact, that for any group there is a 3-regular graph with that group. The smallest 3-regular graph whose group is the identity is the following:



(ii) Folkman (1967) proved that every edge- but not vertex-transitive regular graph has at least 20 vertices. This result is best possible:



The Folkman graph

The *Gray graph* (see Bouwer, 1972) is a 3-regular edge- but not vertex-transitive graph on 54 vertices. It has the following description: take three copies of $K_{3,3}$. For a particular edge e, subdivide e in each of the three

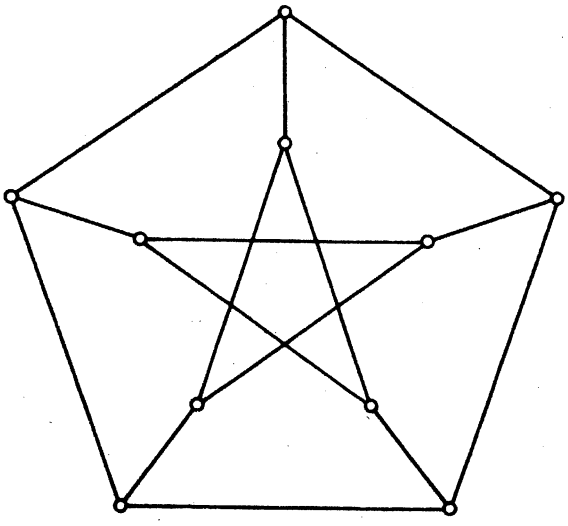
copies and join the resulting three vertices to a new vertex. Repeat this with each edge.

CAGES

An m-regular graph of girth n with the least possible number of vertices is called an (m, n)-*cage*. If we denote by $f(m, n)$ the number of vertices in an (m, n)-cage, it is easy to see that $f(2, n) = n$ and for $m \geq 3$,

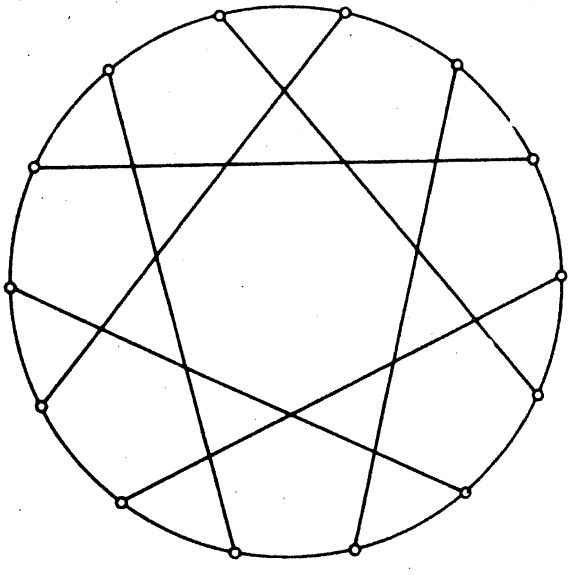
$$f(m, n) \geq \begin{cases} \dfrac{m(m-1)^r - 2}{m-2} & \text{if} \quad n = 2r+1 \\ \dfrac{2(m-1)^r - 2}{m-2} & \text{if} \quad n = 2r \end{cases} \tag{III.1}$$

The $(2, n)$-cage is the n-cycle, the $(m, 3)$-cage is K_{m+1}, and the $(m, 4)$-cage is $K_{m,m}$. In each of these cases, equality holds in (III.1). It has been shown by Hoffman and Singleton (1960) that, for $m \geq 3$ and $n \geq 5$, equality can hold in (III.1) only if $n = 5$ and $m = 3$, 7 or 57, or $n = 6$, 8 or 12. When $m - 1$ is a prime power, the $(m, 6)$-cage is the point-line incidence graph of the projective plane of order $m - 1$; the $(m, 8)$- and $(m, 12)$-cages are also obtained from projective geometries (see Biggs, 1974 for further details). Some of the smaller (m, n)-cages are depicted below:



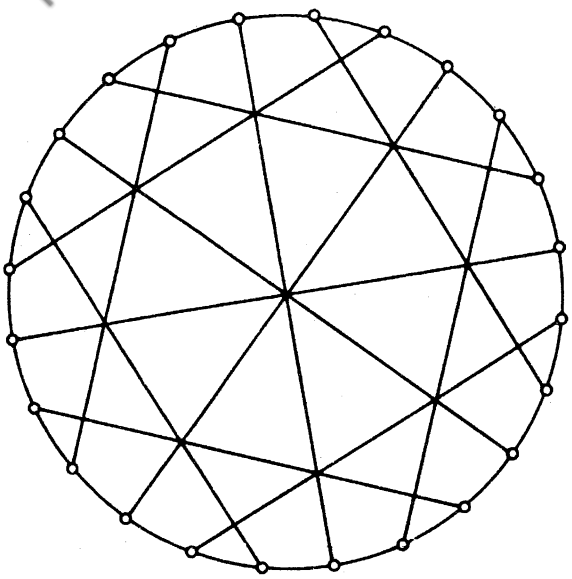
(3,5) - cage

The Petersen graph



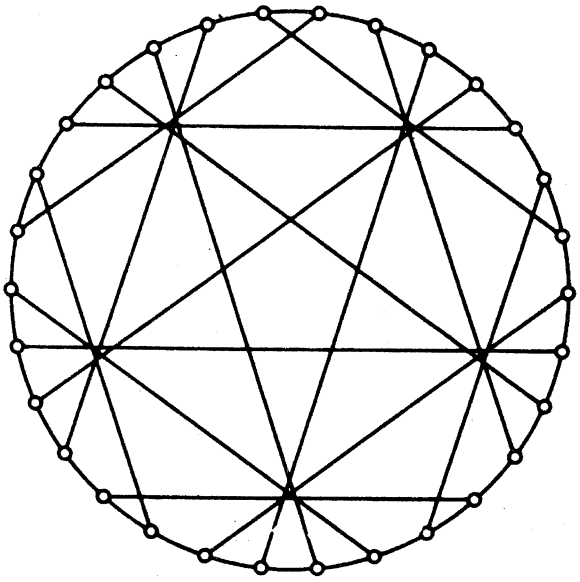
(3,6) - cage

The Heawood graph



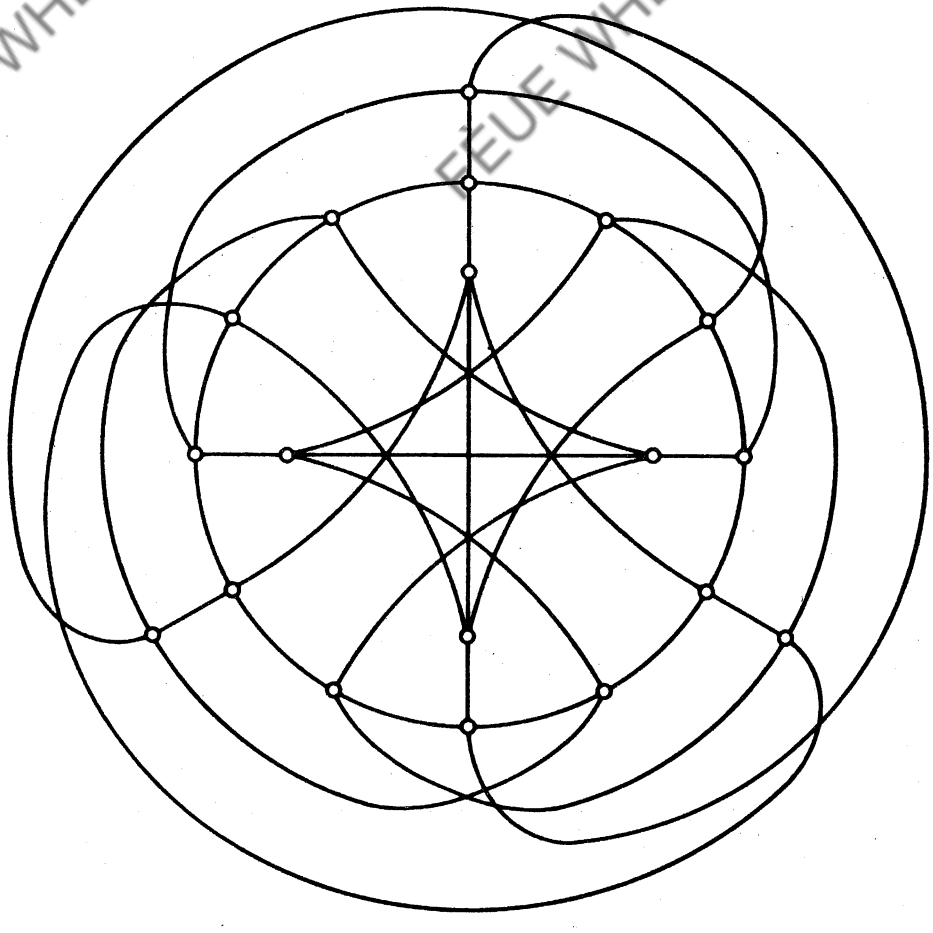
(3,7) – cage

The McGee graph



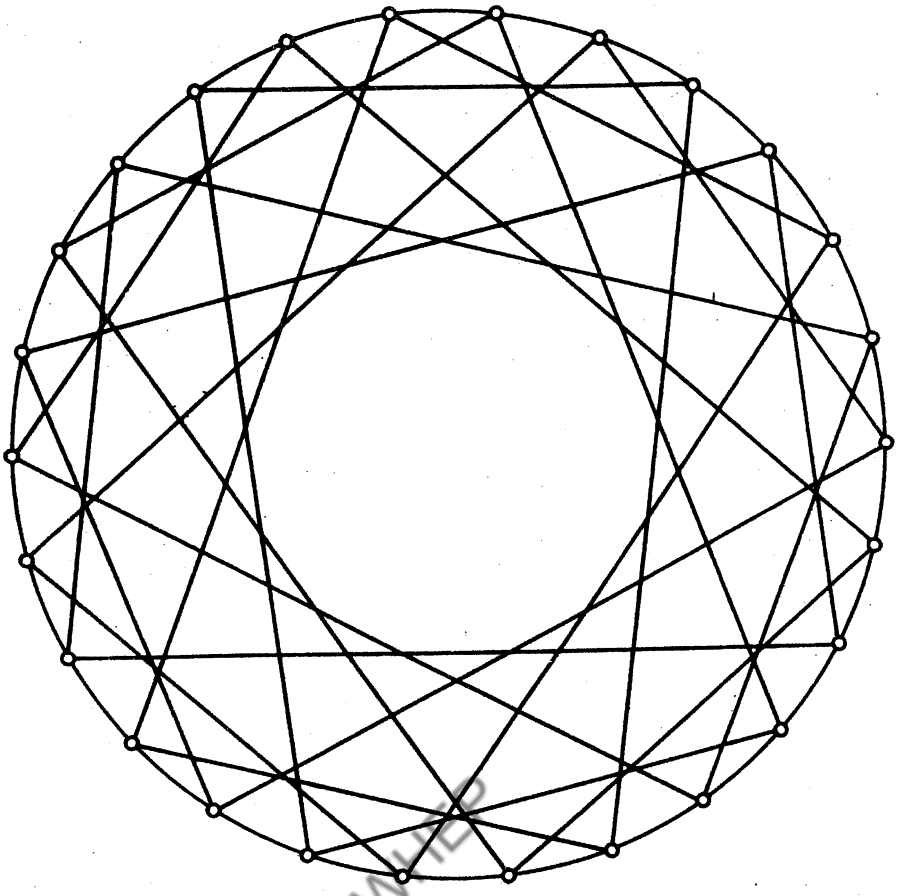
(3,8) – cage

The Tutte–Coxeter graph

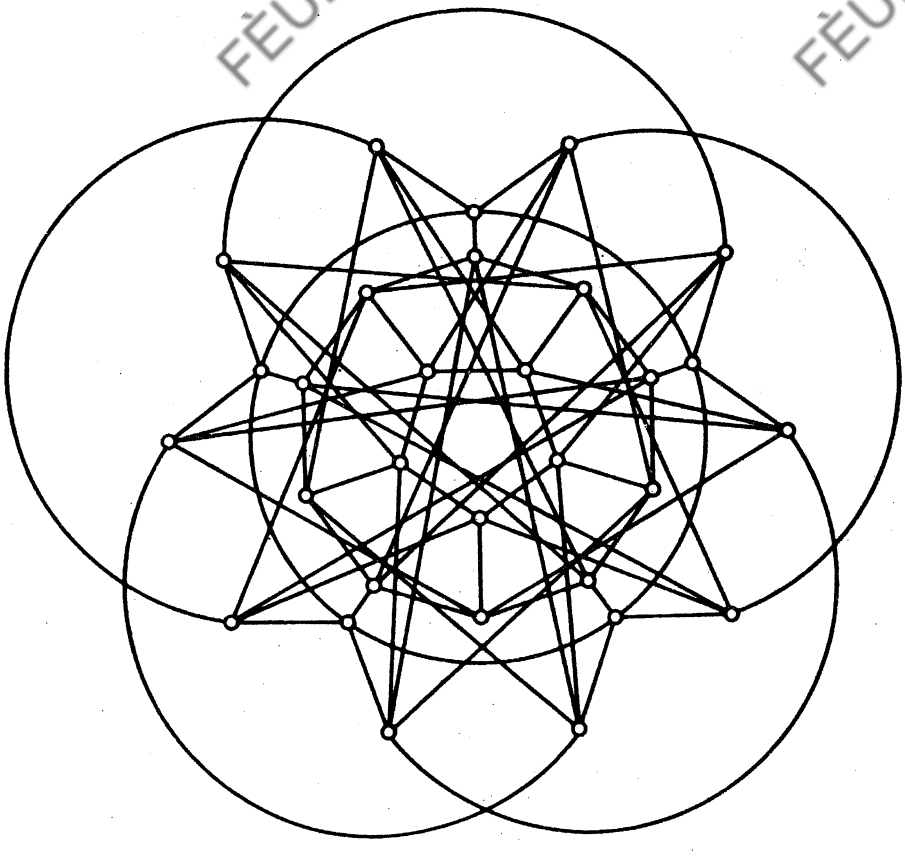


(4,5) – cage

The Robertson graph



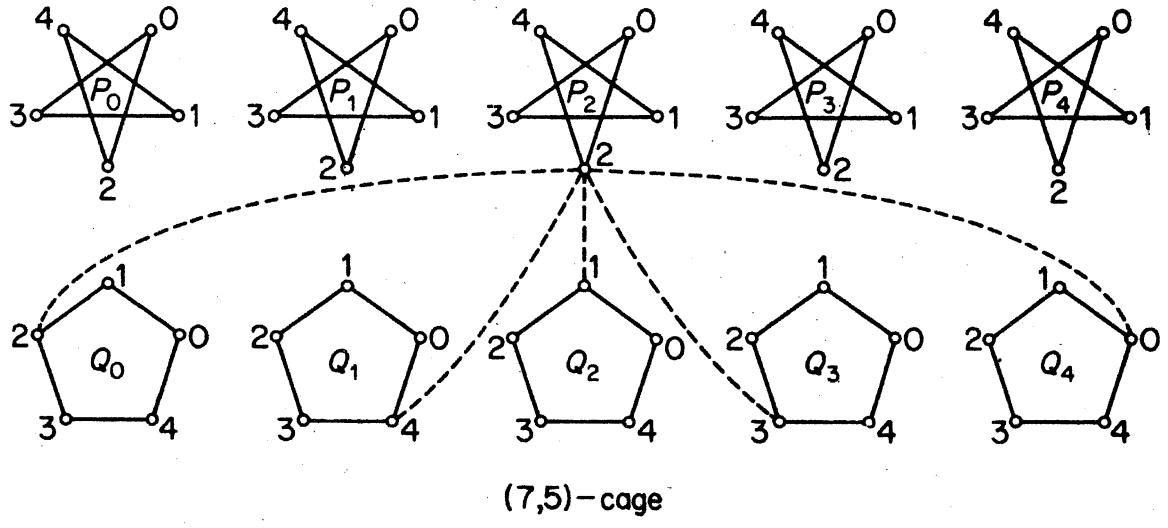
(4,6) – cage



(5,5) – cage

The Robertson–Wegner graph

The (7, 5)-cage (the *Hoffman–Singleton graph*) can be described as follows: it has ten 5-cycles P_0, P_1, P_2, P_3, P_4, Q_0, Q_1, Q_2, Q_3, Q_4, labelled as shown below; vertex i of P_j is joined to vertex $i+jk \pmod 5$ of Q_k. (For example, vertex 2 of P_2 is connected as indicated.)

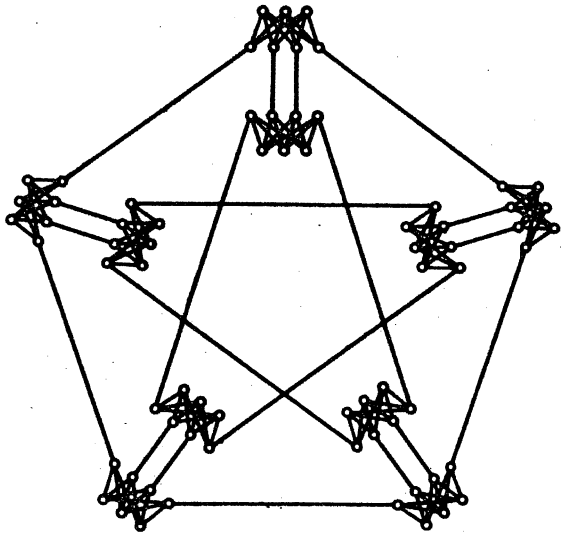


The Hoffman–Singleton graph

NONHAMILTONIAN GRAPHS

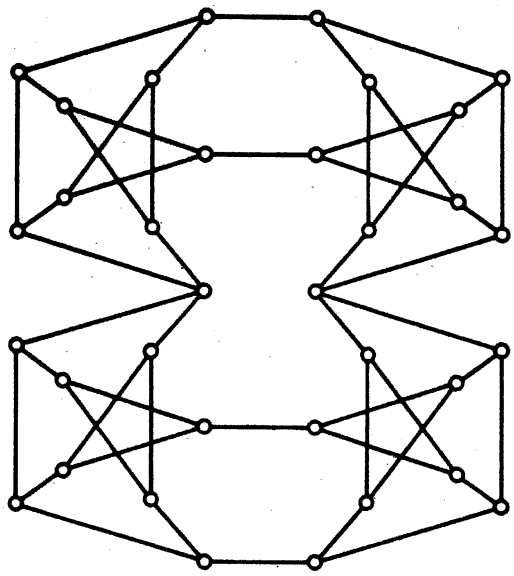
(i) Conditions for a graph to be hamiltonian have been sought ever since Tait made his conjecture on planar graphs. Listed here are counter-examples to several conjectured results.

(a) Every 4-regular 4-connected graph is hamiltonian (C. St. J. A. Nash-Williams).



The Meredith graph

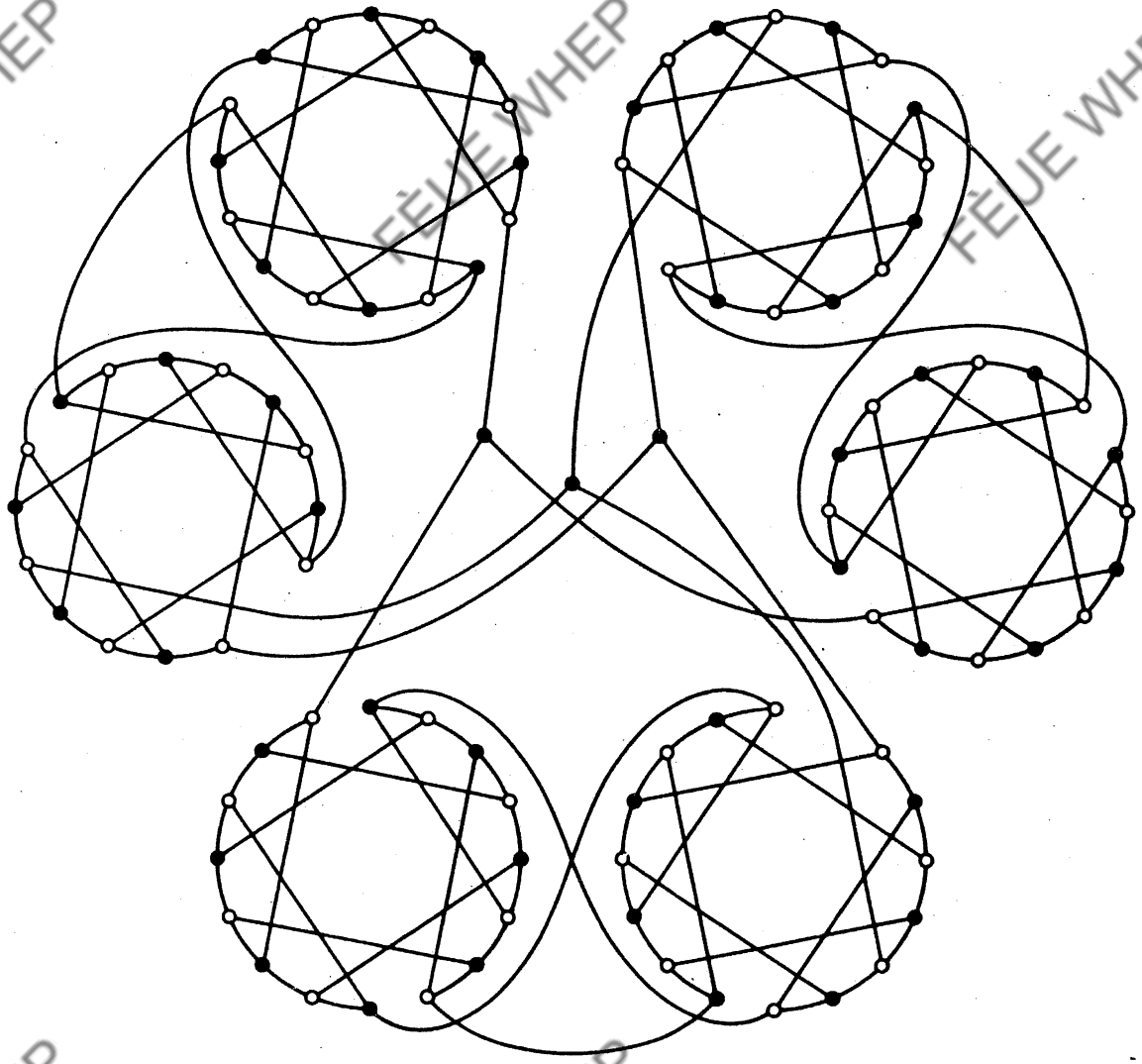
(b) There is no hypotraceable graph (T. Gallai).



The Thomassen graph

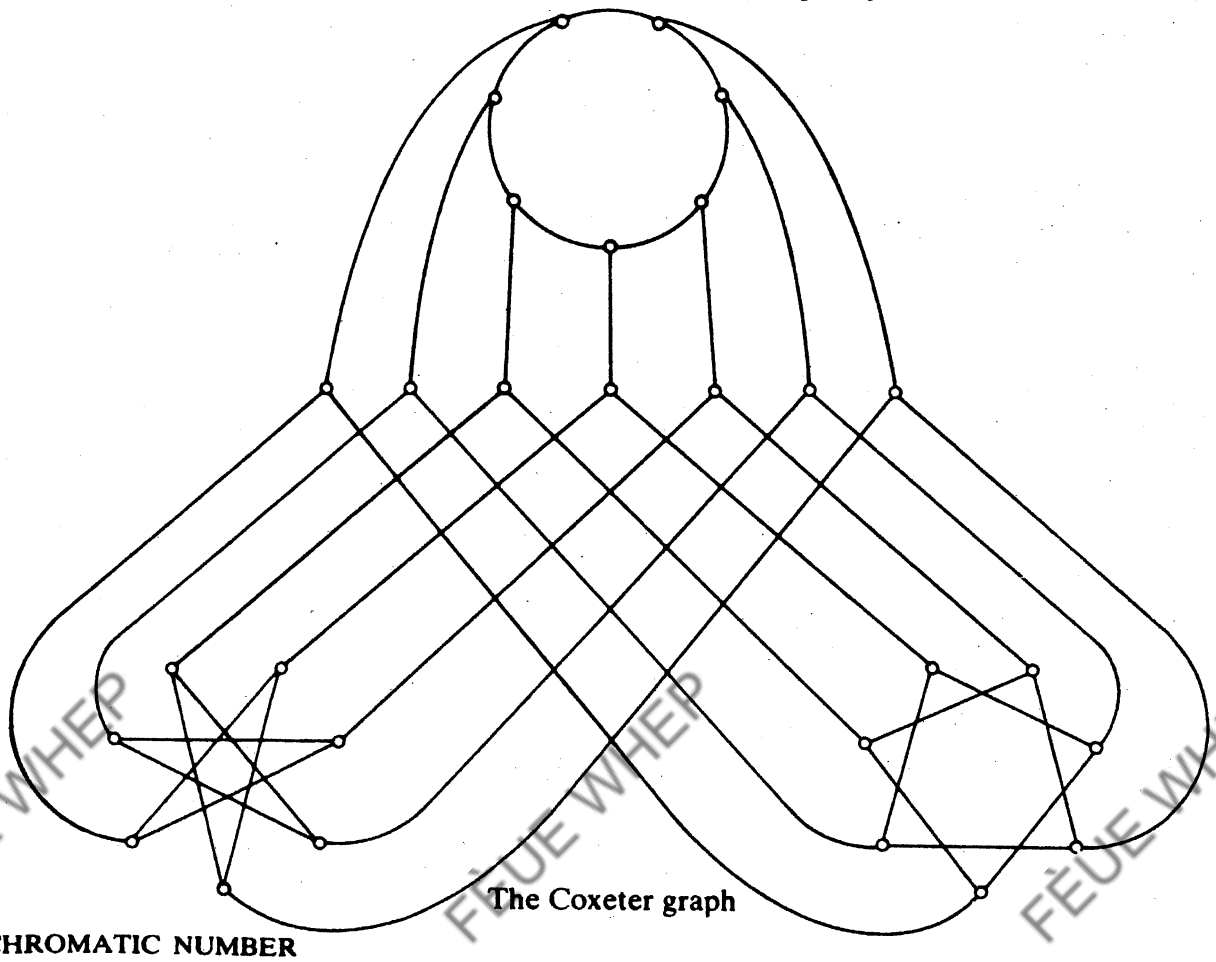
(The first hypotraceable graph was discovered by J. D. Horton.)

(c) Every 3-regular 3-connected bipartite graph is hamiltonian (W. T. Tutte).



The Horton graph

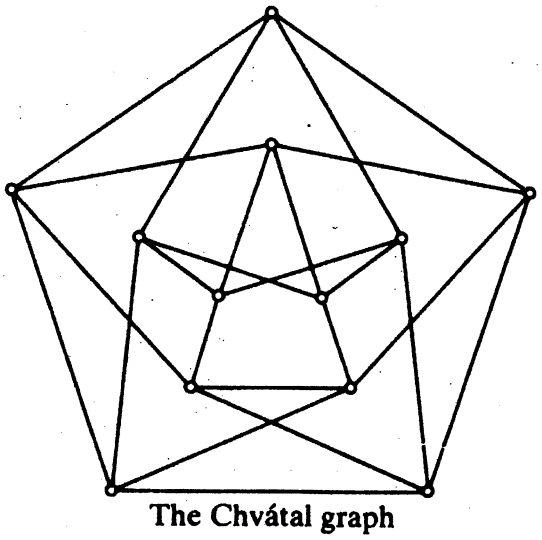
(ii) An example of a nonhamiltonian graph with a high degree of symmetry—there is an automorphism taking any path of length three into any other. (The Petersen graph also has this property.) See Tutte (1960).



The Coxeter graph

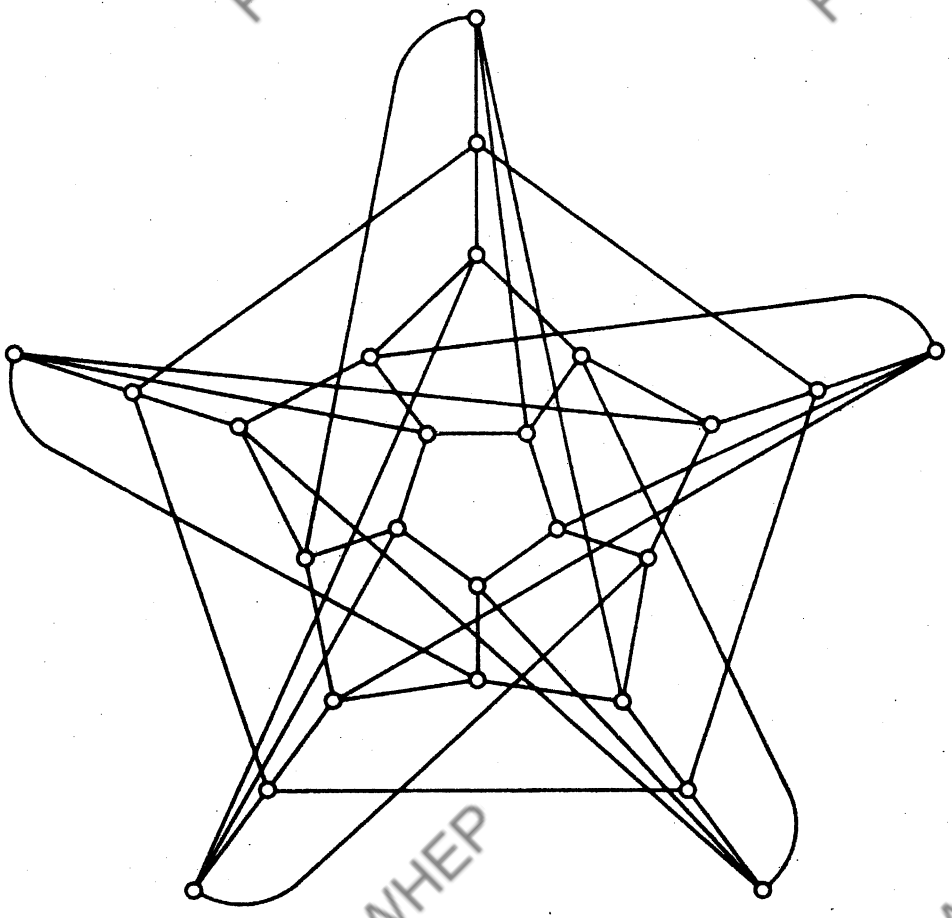
CHROMATIC NUMBER

(i) Grünbaum (1970) has conjectured that, for every $m > 1$ and $n > 2$, there exists an m-regular, m-chromatic graph of girth at least n. For $n = 3$, this is trivial, and for $m = 2$ and 3, the validity of the conjecture follows from the existence of the cages†. Apart from this, only two such graphs are known:



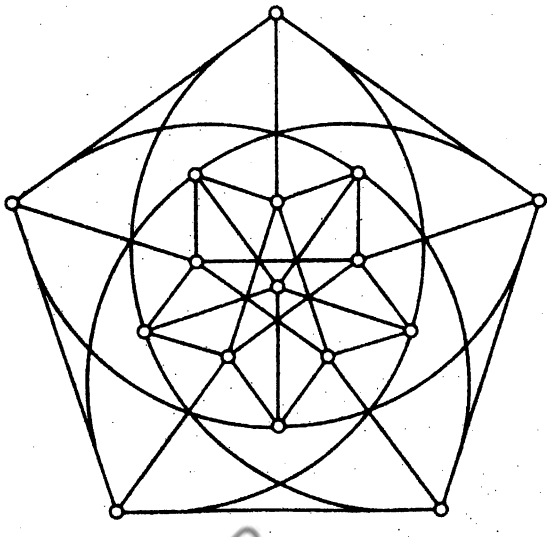
The Chvátal graph

† This conjecture has now been disproved: (Borodin, O. V. and Kostochka, A. V. (1976). On an upper bound of the graph's chromatic number depending on graph's degree and density. *Inst. Maths.*, Novosibirsk, preprint GT-7).



The Grünbaum graph

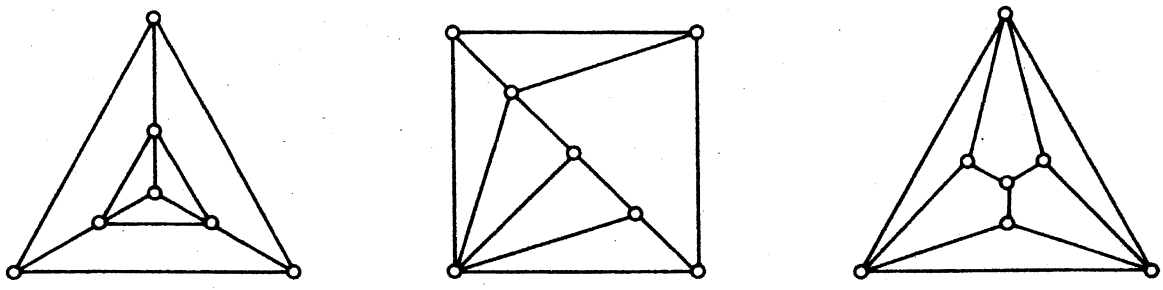
(ii) Since $r(3, 3, 3) = 17$ (see exercise 7.2.3), there is a 3-edge colouring of K_{16} without monochromatic triangles. Kalbfleisch and Stanton (1968) showed that, in such a colouring, the subgraph induced by the edges of any one colour is isomorphic to the following graph:



The Greenwood–Gleason graph

EMBEDDINGS

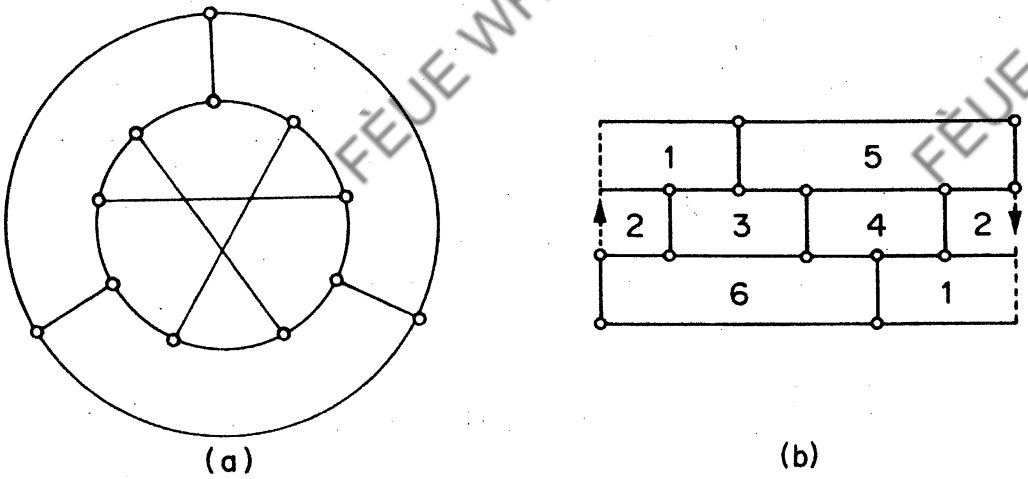
(i) Simple examples of self-dual plane graphs are the wheels. Some more interesting plane graphs with this property are depicted below (see, for example, Smith and Tutte, 1950).



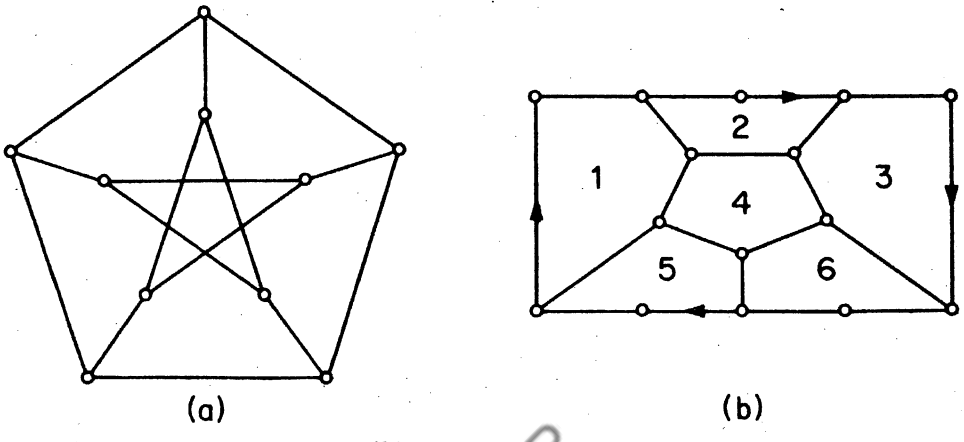
(ii) The *chromatic number* $\chi(S)$ of a surface S is the maximum number of colours required to properly colour the faces of any map on S. (The four-colour conjecture claims that the sphere is 4-chromatic.) Heawood (1890) proved that if S has characteristic $n<2$, then

$$\chi(S)\leq[\tfrac{1}{2}(7+\sqrt{49-24n})] \qquad \text{(III.2)}$$

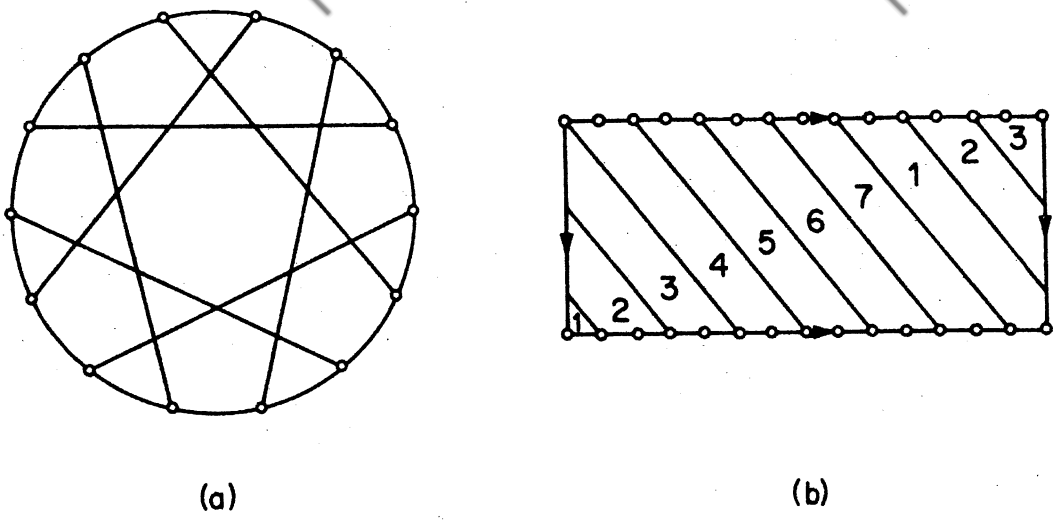
For the projective plane and Möbius band (characteristic 1) and for the torus (characteristic 0), the bound given in (III.2) is attained, as is shown by the following graphs and their embeddings:



(*a*) The Tietze graph; (*b*) an embedding on the Möbius band

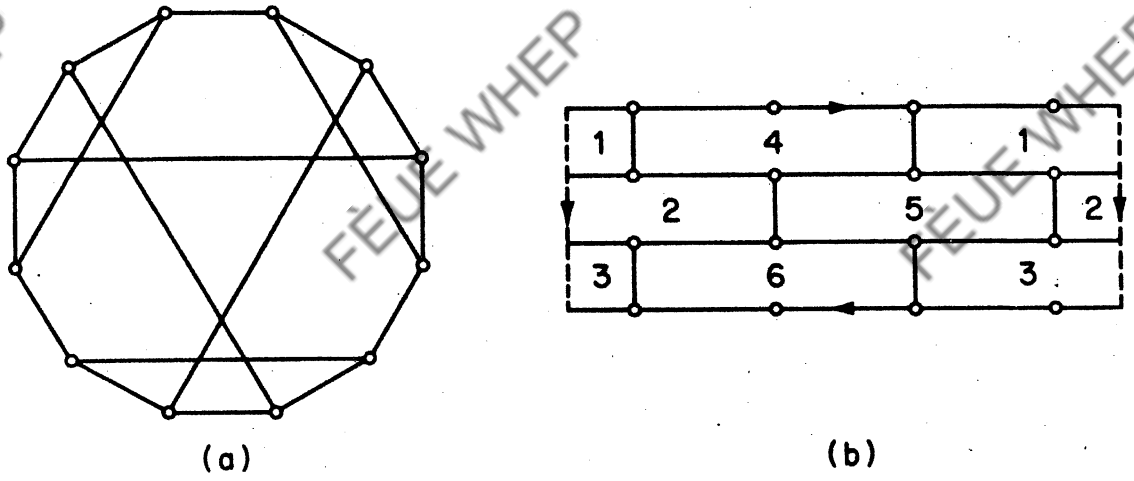


(*a*) The Petersen graph; (*b*) an embedding on the projective plane



(*a*) The Heawood graph; (*b*) an embedding on the torus

Although the Klein bottle has characteristic 0, Franklin (1934) proved that it is only 6-chromatic, and found the following 6-chromatic map on the Klein bottle:



(*a*) The Franklin graph; (*b*) an embedding on the Klein bottle

It has been shown that, with the sole exception of the Klein bottle, equality holds in (III.2) for every surface S of characteristic $n<2$. This result is known as the *map colour theorem* (see Ringel, 1974).

REFERENCES

Biggs, N. (1974). *Algebraic Graph Theory*, Cambridge University Press

Bouwer, I. Z. (1972). On edge but not vertex transitive regular graphs. *J. Combinatorial Theory* B, **12,** 32–40

Coxeter, H. S. M. (1950). Self-dual configurations and regular graphs. *Bull. Amer. Math. Soc.*, **56,** 413–55

Folkman, J. (1967). Regular line-symmetric graphs. *J. Combinatorial Theory*, **3,** 215–32

Franklin, P. (1934). A six color problem. *J. Math. Phys.*, **13,** 363–69

Fréchet, M. and Ky Fan (1967). *Initiation to Combinatorial Topology*, Prindle, Weber and Schmidt, Boston

Frucht, R. (1949). Graphs of degree three with a given abstract group. *Canad. J. Math.*, **1,** 365-78

Grünbaum, B. (1970). A problem in graph coloring. *Amer. Math. Monthly*, **77,** 1088–1092

Heawood, P. J. (1890). Map colour theorem. *Quart. J. Math.*, **24,** 332–38

Hoffman, A. J. and Singleton, R. R. (1960). On Moore graphs with diameters 2 and 3, *IBM J. Res. Develop.*, **4,** 497–504

Horton, J. D. (1974) to be published

Kalbfleisch, J. and Stanton, R. (1968). On the maximal triangle-free edge-chromatic graphs in three colors. *J. Combinatorial Theory*, **5,** 9–20

Meredith, G. H. J. (1973). Regular n-valent n-connected nonhamiltonian non-n-edge-colorable graphs. *J. Combinatorial Theory* B, **14,** 55–60

Ringel, G. (1974). *Map Color Theorem*, Springer-Verlag, Berlin

Smith, C. A. B. and Tutte, W. T. (1950). A class of self-dual maps. *Canad. J. Math.*, **2,** 179–96

Thomassen, C. (1974). Hypohamiltonian and hypotraceable graphs. *Discrete Math.*, **9,** 91–96

Tutte, W. T. (1960). A non-Hamiltonian graph. *Canad. Math. Bull.*, **3,** 1–5

Wegner, G. (1973). A smallest graph of girth 5 and valency 5. *J. Combinatorial Theory* B, **14,** 203–208

Appendix IV
Unsolved Problems

Collected here are a number of unsolved problems of varying difficulty, with originators, dates and relevant bibliography. Conjectures are displayed in bold type. Problems marked † have now been solved; see page 253.

1. Two graphs G and H are *hypomorphic* (written $G \cong H$) if there is a bijection $\sigma: V(G) \to V(H)$ such that $G - v \cong H - \sigma(v)$ for all $v \in V(G)$. A graph G is *reconstructible* if $G \cong H$ implies $G \cong H$. The *reconstruction conjecture* claims that **every graph G with $\nu > 2$ is reconstructible** (S. M. Ulam, 1929). This has been verified for disconnected graphs, trees and a few other classes of graphs (see Harary, 1974).

   There is a corresponding *edge reconstruction conjecture:* **every graph G with $\varepsilon > 3$ is edge reconstructible.** Lovász (1972) has shown that every simple graph G with $\varepsilon > \binom{\nu}{2} \Big/ 2$ is edge reconstructible.

   P. K. Stockmeyer has found an infinite family of counterexamples to the analogous reconstruction conjecture for digraphs.

   Bondy, J. A. and Hemminger, R. L. (1976). Graph reconstruction—a survey. *J. Graph Theory*, to be published

   Lovász, L. (1972). A note on the line reconstruction problem. *J. Combinatorial Theory* B, **13,** 309–10

2. A graph G is *embeddable* in a graph H if G is isomorphic to a subgraph of H. Characterise the graphs embeddable in the k-cube (V. V. Firsov, 1965).

   Garey, M. R. and Graham, R. L. (1975). On cubical graphs. *J. Combinatorial Theory* (B), **18,** 84–95

3. **Every 4-regular simple graph contains a 3-regular subgraph** (N. Sauer, 1973).

4. **If $k > 2$, there exists no graph with the property that every pair of vertices is connected by a unique path of length k** (A. Kotzig, 1974). Kotzig has verified his conjecture for $k < 9$.

5. **Every connected graph G is the union of at most $[(\nu+1)/2]$ edge-disjoint paths** (T. Gallai, 1962). Lovász (1968) has shown that every graph G is the union of at most $[\nu/2]$ edge-disjoint paths and cycles.

Lovász, L. (1968). On coverings of graphs, in *Theory of Graphs* (eds. P. Erdös and G. Katona), Academic Press, New York, pp. 231–36

6. **Every 2-edge-connected simple graph G is the union of $\nu-1$ cycles** (P. Erdös, A. W. Goodman and L. Pósa, 1966).

Erdös, P., Goodman, A. W. and Pósa, L. (1966). The representation of a graph by set intersections. *Canad. J. Math.*, **18,** 106–12

7. **If G is a simple block with at least $\nu/2+k$ vertices of degree at least k, then G has a cycle of length at least $2k$** (D. R. Woodall, 1975).
8. Let $f(m, n)$ be the maximum possible number of edges in a simple graph on n vertices which contains no m-cycle. It is known that

$$f(m,n)=\begin{cases}[n^2/4] & \text{if } m \text{ is odd}, \quad m\le\tfrac{1}{2}(n+3)\\ \binom{n-m+2}{2}+\binom{m-1}{2} & \text{if } m\ge\tfrac{1}{2}(n+3)\end{cases}$$

Determine $f(m, n)$ for the remaining cases (P. Erdös, 1963).

Bondy, J. A. and Simonovits, M. (1974). Cycles of even length in graphs. *J. Combinatorial Theory* (B), **16,** 97–105

Woodall, D. R. (1972). Sufficient conditions for circuits in graphs. *Proc. London Math. Soc.*, **24,** 739–55

9. Let $f(n)$ be the maximum possible number of edges in a simple graph on n vertices which contains no 3-regular subgraph. Determine $f(n)$ (P. Erdös and N. Sauer, 1974). Since there is a constant c such that every simple graph G with $\varepsilon\ge c\nu^{8/5}$ contains the 3-cube (Erdös and Simonovits, 1970), clearly $f(n)<cn^{8/5}$.

Erdös, P. and Simonovits, M. (1970). Some extremal problems in graph theory, in *Combinatorial Theory and its Applications I* (eds. P. Erdös, A. Rényi and V. T. Sós), North-Holland, Amsterdam, pp. 378–92

10. Determine which simple graphs G have exactly one cycle of each length l, $3\le l\le\nu$ (R. C. Entringer, 1973).
11. Let $f(n)$ be the maximum possible number of edges in a graph on n vertices in which no two cycles have the same length. Determine $f(n)$ (P. Erdös, 1975).
12. **If G is simple and $\varepsilon>\nu(k-1)/2$, then G contains every tree with k edges** (P. Erdös and V. T. Sós, 1963). It is known that every such graph contains a path of length k (Erdös and Gallai, 1959).

Erdös, P. and Gallai, T. (1959). On maximal paths and circuits of graphs. *Acta Math. Acad. Sci. Hungar.*, **10,** 337–56

13. Find a (6, 5)-cage (see appendix III).

14. The *bandwidth* of G is defined to be

$$\min_{l} \max_{uv \in E} |l(u) - l(v)|$$

where the minimum is taken over all labellings l of V in distinct integers. Find bounds for the bandwidth of a graph (L. H. Harper, 1964). The bandwidth of the k-cube has been determined by Harper (1966).

Chvátalová, J. (1975). Optimal labelling of a product of two paths. *Discrete Math.*, **11,** 249–53

Harper, L. H. (1966). Optimal numberings and isoperimetric problems on graphs. *J. Combinatorial Theory*, **1,** 385–93

15. A simple graph G is *graceful* if there is a labelling l of its vertices with distinct integers from the set $\{0, 1, \ldots, \varepsilon\}$, so that the induced edge labelling l' defined by

$$l'(uv) = |l(u) - l(v)|$$

assigns each edge a different label. Characterise the graceful graphs (S. Golomb, 1972). It has been conjectured that, in particular, **every tree is graceful** (A. Rosa, 1966).

Golomb, S. (1972). How to number a graph, in *Graph Theory and Computing* (ed. R. C. Read), Academic Press, New York, pp. 23–37

†16. **The 3-connected planar graph on $2m$ edges with the least possible number of spanning trees is the wheel with m spokes** (W. T. Tutte, 1940).

Kelmans, A. K. and Chelnokov, V. M. (1974). A certain polynomial of a graph and graphs with an extremal number of trees. *J. Combinatorial Theory* (B), **16,** 197–214

17. Let u and v be two vertices in a graph G. Denote the minimum number of vertices whose deletion destroys all (u, v)-paths of length at most n by a_n, and the maximum number of internally disjoint (u, v)-paths of length at most n by b_n. Let $f(n)$ denote the maximum possible value of a_n/b_n. Determine $f(n)$ (V. Neumann, 1974). L. Lovász has conjectured that $f(n) \leq \sqrt{n}$. It is known that

$$[\sqrt{n/2}] \leq f(n) \leq [n/2]$$

18. **Every 3-regular 3-connected bipartite planar graph is hamiltonian** (D. Barnette, 1970). P. Goodey has verified this conjecture for plane graphs whose faces are all of degree four or six. Note that if the planarity condition is dropped, the conjecture is no longer valid (see appendix III).

19. A graphic sequence $\mathbf{d}$ is *forcibly hamiltonian* if every simple graph with degree sequence $\mathbf{d}$ is hamiltonian. Characterise the forcibly hamiltonian

sequences (C. St. J. A. Nash-Williams, 1970). (Theorem 4.5 gives a partial solution.)

Nash-Williams, C. St. J. A. (1970). Valency sequences which force graphs to have Hamiltonian circuits: interim report, University of Waterloo preprint

20. **Every connected vertex-transitive graph has a Hamilton path** (L. Lovász, 1968). L. Babai has verified this conjecture for graphs with a prime number of vertices.

21. A graph G is *t-tough* if, for every vertex cut S, $\omega(G-S) \leq |S|/t$. (Thus theorem 4.2 says that every hamiltonian graph is 1-tough.)

(*a*) **If G is 2-tough, then G is hamiltonian** (V. Chvátal, 1971). C. Thomassen has obtained an example of a nonhamiltonian t-tough graph with $t > 3/2$

(*b*) **If G is 3/2-tough, then G has a 2-factor** (V. Chvátal, 1971).

Chvátal, V. (1973). Tough graphs and hamiltonian circuits. *Discrete Math.*, **5,** 215–28

22. The *binding number* of G is defined by

$$\text{bind } G = \min_{\substack{\emptyset \neq S \subseteq V \\ N(S) \neq V}} |N(S)|/|S|$$

(*a*) **If bind $G \geq 3/2$, then G contains a triangle** (D. R. Woodall, 1973).

(*b*) **If bind $G \geq 3/2$, then G is pancyclic** (contains cycles of all lengths l, $3 \leq l \leq v$) (D. R. Woodall, 1973).

Woodall (1973) has shown that G is hamiltonian if bind $G \geq 3/2$, and that G contains a triangle if bind $G \geq \frac{1}{2}(1+\sqrt{5})$.

Woodall, D. R. (1973). The binding number of a graph and its Anderson number. *J. Combinatorial Theory* (B), **15,** 225–55

23. **Every nonempty regular simple graph contains two disjoint maximal independent sets** (C. Payan, 1973)

24. Find the Ramsey number $r(3, 3, 3, 3)$. It is known that

$$51 \leq r(3, 3, 3, 3) \leq 65$$

Chung, F. R. K. (1973). On the Ramsey numbers $N(3, 3, \ldots, 3; 2)$, *Discrete Math.*, **5,** 317–21

Folkman, J. (1974). Notes on the Ramsey number $N(3, 3, 3, 3)$. *J. Combinatorial Theory* (A), **16,** 371–79

25. For $m < n$, let $f(m, n)$ denote the least possible number of vertices in a graph which contains no K_n but has the property that in every 2-edge colouring there is a monochromatic K_m. (Folkman, 1970 has established the existence of such graphs.) Determine bounds for $f(m, n)$. It is

known that

$$f(3, n) = 6 \quad \text{for} \quad n \geq 7$$
$$f(3, 6) = 8 \quad \text{(see exercise 7.2.5)}$$
$$10 \leq f(3, 5) \leq 18$$

Folkman, J. (1970). Graphs with monochromatic complete subgraphs in every edge coloring. *SIAM J. Appl. Math.*, **18,** 19–24

Irving, R. W. (1973). On a bound of Graham and Spencer for a graph-colouring constant. *J. Combinatorial Theory* (B), **15,** 200–203

Lin, S. On Ramsey numbers and K_r-coloring of graphs. *J. Combinatorial Theory* (B), **12,** 82–92

26. **If G is n-chromatic, then $r(G, G) \geq r(n, n)$** (P. Erdös, 1973). ($r(G, G)$ is defined in exercise 7.2.6.)
27. What is the maximum possible chromatic number of a graph which can be drawn in the plane so that each edge is a straight line segment of unit length? (L. Moser, 1958).

Erdös. P., Harary, F. and Tutte, W. T. (1965). On the dimension of a graph. *Mathematika*, **12,** 118–22

28. **The absolute values of the coefficients of any chromatic polynomial form a unimodal sequence** (that is, no term is flanked by terms of greater value) (R. C. Read, 1968).

Chvátal, V. (1970). A note on coefficients of chromatic polynomials. *J. Combinatorial Theory*, **9,** 95–96

29. **If G is not complete and $\chi = m + n - 1$, where $m \geq 2$ and $n \geq 2$, then there exist disjoint subgraphs G_1 and G_2 of G such that $\chi(G_1) = m$ and $\chi(G_2) = n$** (L. Lovász, 1968).
30. A simple graph G is *perfect* if, for every induced subgraph H of G, the number of vertices in a maximum clique is $\chi(H)$. **G is perfect if and only if no induced subgraph of G or G^c is an odd cycle of length greater than three** (C. Berge, 1961). This is the *strong perfect graph conjecture.* Lovász (1972) has shown that the complement of any perfect graph is perfect.

Lovász, L. (1972). Normal hypergraphs and the perfect graph conjecture. *Discrete Math.*, **2,** 253–67

Parthasarathy, K. R. and Ravindra, G. (to be published). The strong perfect-graph conjecture is true for $K_{1,3}$-free graphs. *J. Combinatorial Theory*

31. **If G is a 3-regular simple block and H is obtained from G by duplicating each edge, then $\chi'(H) = 6$** (D. R. Fulkerson, 1971).
32. **If G is simple, with ν even and $\chi'(G) = \Delta(G) + 1$, then $\chi'(G - v) = \chi'(G)$**

for some $v \in V$ (I. T. Jakobsen, L. W. Beineke and R. J. Wilson, 1973). This has been verified for all graphs G with $\nu \leq 10$ and all 3-regular graphs G with $\nu = 12$.

Beineke, L. W. and Wilson, R. J. (1973). On the edge-chromatic number of a graph. *Discrete Math.*, **5,** 15–20

33. **For any simple graph G, the elements of $V \cup E$ can be coloured in $\Delta+2$ colours so that no two adjacent or incident elements receive the same colour** (M. Behzad, 1965). This is known as the *total colouring conjecture.* M. Rosenfeld and N. Vijayaditya have verified it for all graphs G with $\Delta \leq 3$.

Vijayaditya, N. (1971). On total chromatic number of a graph. *J. London Math. Soc.*, **3,** 405–408

34. **If G is simple and $\varepsilon > 3\nu - 6$, then G contains a subdivision of K_5** (G. A. Dirac, 1964). Thomassen (1975) has shown that G contains a subdivision of K_5 if $\varepsilon \geq 4\nu - 10$.

Dirac, G. A. (1964). Homomorphism theorems for graphs. *Math. Ann.*, **153,** 69–80

Thomassen, C. (1974). Some homeomorphism properties of graphs, *Math. Nachr.*, **64,** 119–33

35. A sequence **d** of non-negative integers is *potentially planar* if there is a simple planar graph with degree sequence **d**. Characterise the potentially planar sequences (S. L. Hakimi, 1963).

Owens, A. B. (1971). On the planarity of regular incidence sequences. *J. Combinatorial Theory* (B), **11,** 201–12

†36. **If G is a loopless planar graph, then $\alpha \geq \nu/4$** (P. Erdös, 1968). Albertson (1974) has shown that every such graph satisfies $\alpha > 2\nu/9$.

Albertson, M. O. (1974). Finding an independent set in a planar graph, in *Graphs and Combinatorics* (eds. R. A. Bari and F. Harary), Springer-Verlag, New York, pp. 173–79

†37. **Every planar graph is 4-colourable** (F. Guthrie, 1852).

Ore, O. (1969). *The Four-Color Problem*, Academic Press, New York

38. **Every k-chromatic graph contains a subgraph contractible to K_k** (H. Hadwiger, 1943). Dirac (1964) has proved that every 6-chromatic graph contains a subgraph contractible to K_6 less one edge.

Dirac, G. A. (1964). Generalizations of the five colour theorem, in *Theory of Graphs and its Applications* (ed. M. Fiedler), Academic Press, New York, pp. 21–27

39. **Every k-chromatic graph contains a subdivision of K_k** (G. Hajós, 1961). Pelikán (1969) has shown that every 5-chromatic graph contains a subdivision of K_5 less one edge.

Pelikán, J. (1969). Valency conditions for the existence of certain subgraphs, in *Theory of Graphs* (eds. P. Erdös and G. Katona), Academic Press, New York, pp. 251–58

40. **Every 2-edge-connected 3-regular simple graph which has no Tait colouring contains a subgraph contractible to the Petersen graph** (W. T. Tutte, 1966).

Isaacs, R. (1975). Infinite families of nontrivial trivalent graphs which are not Tait colourable. *Amer. Math. Monthly*, **82**, 221–39

Tutte, W. T. (1966). On the algebraic theory of graph colorings. *J. Combinatorial Theory*, **1**, 15–50

41. **For every surface S, there exists a finite number of graphs which have minimum degree at least three and are minimally nonembeddable on S.**

†42. **If D is diconnected, then D has a directed cycle of length at least χ** (M. Las Vergnas, 1974).

43. **If D is a tournament with v odd and every indegree and outdegree equal to $(v-1)/2$, then D is the union of $(v-1)/2$ arc-disjoint directed Hamilton cycles** (P. Kelly, 1966).

44. **If D is a tournament with v even, then D is the union of $\sum_{v \in V} \max\{0, d^+(v) - d^-(v)\}$ arc-disjoint directed paths** (R. O'Brien, 1974).

This would imply the truth of conjecture 43.

45. Characterise the tournaments D with the property that all subtournaments $D - v$ are isomorphic (A. Kotzig, 1973).

46. **If D is a digraph which contains a directed cycle, then there is some arc whose reversal decreases the number of directed cycles in D** (A. Adám, 1963).

47. **Given a positive integer n, there exists a least integer $f(n)$ such that in any digraph with at most n arc-disjoint directed cycles there are $f(n)$ arcs whose deletion destroys all directed cycles** (T. Gallai, 1964; D. H. Younger, 1968).

Erdös. P. and Pósa, L. (1962). On the maximal number of disjoint circuits of a graph. *Publ. Math. Debrecen*, **9**, 3–12

Younger, D. H. (1973). Graphs with interlinked directed circuits, in *Proceedings of Midwest Symposium on Circuit Theory*

48. An $(m+n)$-regular graph is (m, n)-*orientable* if it can be oriented so that each indegree is either m or n. **Every 5-regular simple graph with no 1-edge cut or 3-edge cut is (4, 1)-orientable** (W. T. Tutte, 1972). Tutte has shown that this would imply Grötzsch's theorem

49. Obtain an algorithm to find a maximum flow in a network with two sources x_1 and x_2, two sinks y_1 and y_2, and two commodities, the requirement being to ship commodity 1 from x_1 to y_1 and commodity 2 from x_2 to y_2 (L. R. Ford and D. R. Fulkerson, 1962).

Rothschild, B. and Whinston, A. (1966). On two commodity network flows. *Operations Res.*, **14,** 377–87

50. **Every 2-edge-connected digraph D has a circulation f over the field of integers modulo 5 in which $f(a) \neq 0$ for all arcs a** (W. T. Tutte, 1949). Tutte has shown that this would imply the five-colour theorem.

References for problems solved since first printing:

16. Göbel, F. and Jagers, A. A. (1976). On a conjecture of Tutte concerning minimal tree numbers. *J. Combinatorial Theory* (B), to be published

36 and 37. Appel, K. and Haken, W. (1976). Every planar map is four colorable. *Bull. Amer. Math. Soc.*, **82,** 711–2

42. Bondy, J. A. (1976). Diconnected orientations and a conjecture of Las Vergnas. *J. London Math. Soc.*, to be published

Appendix V Suggestions for Further Reading

BOOKS OF A GENERAL NATURE, LISTED ACCORDING TO LEVEL OF TREATMENT

Ore, O. (1963). *Graphs and Their Uses*, Random House, New York
Rouse Ball, W. W. and Coxeter, H. S. M. (1974). *Mathematical Recreations and Essays*, University of Toronto Press, Toronto
Liu, C. L. (1968). *Introduction to Combinatorial Mathematics*, McGraw-Hill, New York
Wilson, R. J. (1972). *Introduction to Graph Theory*, Oliver and Boyd, Edinburgh
Deo, N. (1974). *Graph Theory with Applications to Engineering and Computer Science*, Prentice-Hall, Englewood Cliffs, N.J.
Behzad, M. and Chartrand, G. (1971). *Introduction to the Theory of Graphs*, Allyn and Bacon, Boston
Harary, F. (ed.) (1967). *A Seminar on Graph Theory*, Holt, Rinehart and Winston, New York
Ore, O. (1962). *Theory of Graphs*, American Mathematical Society, Providence, R.I.
König, D. (1950). *Theorie der Endlichen und Unendlichen Graphen*, Chelsea, New York
Sachs, H. (1970). *Einführung in die Theorie der Endlichen Graphen*, Teubner Verlagsgesellschaft, Leipzig
Harary, F. (1969). *Graph Theory*, Addison-Wesley, Reading, Mass.
Berge, C. (1973). *Graphs and Hypergraphs*, North Holland, Amsterdam

SPECIAL TOPICS

Biggs, N. (1974). *Algebraic Graph Theory*, Cambridge University Press, Cambridge
Tutte, W. T. (1966). *Connectivity in Graphs*, University of Toronto Press, Toronto
Ore, O. (1967). *The Four-Color Problem*, Academic Press, New York
Ringel, G. (1974). *Map Color Theorem*, Springer-Verlag, Berlin

Moon, J. W. (1968). *Topics on Tournaments*, Holt, Rinehart and Winston, New York

Ford, L. R. Jr. and Fulkerson, D. R. (1962). *Flows in Networks*, Princeton University Press, Princeton

Berge, C. and Ghouila-Houri, A. (1965). *Programming, Games, and Transportation Networks*, John Wiley, New York

Seshu, S. and Reed, M. B. (1961). *Linear Graphs and Electrical Networks*, Addison-Wesley, Reading, Mass.

Tutte, W. T. (1971). *Introduction to the Theory of Matroids*, American Elsevier, New York

Harary, F. and Palmer, E. (1973). *Graphical Enumeration*, Academic Press, New York

Aho, A. V., Hopcroft, J. E. and Ullman, J. D. (1974). *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, Mass.

Welsh, D. J. A. (1976). *Matroid Theory*, Academic Press, New York

Biggs, N., Lloyd, E. K. and Wilson, R. J. (1976). *Graph Theory 1736–1936*, Clarendon Press, Oxford

Glossary of Symbols

GENERAL MATHEMATICAL SYMBOLS

		Page
$\cup$	union	
$\cap$	intersection	
$\subseteq$	subset	
$\subset$	proper subset	
$\setminus$	set-theoretic difference	
Δ	symmetric difference	
$[x]$	greatest integer $\leq x$	
$\{x\}$	least integer $\geq x$	
$\|\|f\|\|$	support of f	215
$\mathbf{R} \mid S$	restriction of $\mathbf{R}$ to S	215
$\mathbf{R}'$	transpose of $\mathbf{R}$	

GRAPH-THEORETIC SYMBOLS

$\mathbf{A}$	arc set	171
$\mathbf{A}$	adjacency matrix of a graph	7
$\boldsymbol{A}$	adjacency matrix of a digraph	173
$b(f)$	boundary of f	140
$\mathscr{B}$	bond space	213
$c(G)$	closure of G	56
cap K	capacity of cut K	194
$\mathscr{C}$	cycle space	212
$d_G(v)$	degree of vertex v in G	10
$d_G(f)$	degree of face f in G	140
$d_D^-(v)$	indegree of v in D	172
$d_D^+(v)$	outdegree of v in D	172
$d_G(u, v)$	distance between u and v in G	14
D	directed graph	171
$D(G)$	associated digraph of G	179
ext J	exterior of J	135
Ext J	closure of ext J	135
E	edge set	1
$f^-(S)$	flow into S	191
$f^+(S)$	flow out of S	191
F	face set	139
$F(B, \tilde{H})$	set of faces of $\tilde{H}$ in which B is drawable	164

		Page
G	graph	1
$G[S]$	subgraph of G induced by S	9
int J	interior of J	135
Int J	closure of int J	135
K_n	complete graph	4
$K_{m,n}$	complete bipartite graph	5
$\mathbf{M}$	incidence matrix of a graph	7
$\boldsymbol{M}$	incidence matrix of a digraph	214
N	network	191
$N_G(S)$	neighbour set of S in G	72
$N_D^-(v)$	in-neighbour set of v in D	175
$N_D^+(v)$	out-neighbour set of v in D	175
$r(k, l)$	Ramsey number	103
$r(k_1, k_2, \ldots, k_m)$	Ramsey number	108
r_n	$r(3, 3, \ldots, 3)$	108
val f	value of flow f	192
V	vertex set	1
$V(B, H)$	set of vertices of attachment of B to H	146
α	independence number	101
α'	edge independence number	102
β	covering number	101
β'	edge covering number	102
δ	minimum degree	10
δ^-	minimum indegree	172
δ^+	minimum outdegree	172
Δ	maximum degree	10
Δ^-	maximum indegree	172
Δ^+	maximum outdegree	172
ε	number of edges	3
κ	connectivity	42
κ'	edge connectivity	42
ν	number of vertices	3
o	number of odd components	76
π_k	chromatic polynomial	125
τ	number of spanning trees	32
ϕ	number of faces	139
χ	chromatic number	117
χ'	edge chromatic number	91
χ^*	face chromatic number	158
ω	number of components	13
$\bar{D}$	converse of D	173
$\hat{D}$	condensation of D	173

		Page
G^c	complement of G	6
G^*	dual of G	140
$\tilde{G}$	planar embedding of G	135
W^{-1}	reverse of walk W	12
$G \cdot e$	contraction of e	32
$G - e$	deletion of e	9
$G + e$	addition of e	9
$G - v$	deletion of v	9
$G + E'$	addition of E'	9
$G - S$	deletion of S	9
$G \cong H$	isomorphism	4
$H \subseteq G$	subgraph	8
$H \subset G$	proper subgraph	8
$G \cup H$	union	9
$G \cap H$	intersection	10
$G + H$	disjoint union	10
$G \times H$	product	96
$G \vee H$	join	58
$\bar{H}(G)$	complement of H in G	29
$[S, T]$	set of edges between S and T	29
(S, T)	set of arcs from S to T	176
WW'	concatenation of walks	12

Index

This index is arranged strictly in alphabetical order according to the first significant word. Thus, 'edge connectivity' is listed under E and 'k-chromatic graph' under C.

W.A. Coppel

UNIVERSITEXT

Number Theory

An Introduction to Mathematics

Second Edition

Springer

Universitext

For other titles in this series, go to
www.springer.com/series/223

W.A. Coppel

Number Theory

An Introduction to Mathematics

Second Edition

Springer

W.A. Coppel
3 Jansz Crescent
2603 Griffith
Australia

ISBN 978-0-387-89485-0 e-ISBN 978-0-387-89486-7
DOI 10.1007/978-0-387-89486-7
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009931687

Mathematics Subject Classification (2000): 11-xx, 05B20, 33E05

For Jonathan, Nicholas, Philip and Stephen

Contents

Part B

Preface to the Second Edition

Undergraduate courses in mathematics are commonly of two types. On the one hand there are courses in subjects, such as linear algebra or real analysis, with which it is considered that every student of mathematics should be acquainted. On the other hand there are courses given by lecturers in their own areas of specialization, which are intended to serve as a preparation for research. There are, I believe, several reasons why students need more than this.

First, although the vast extent of mathematics today makes it impossible for any individual to have a deep knowledge of more than a small part, it is important to have some understanding and appreciation of the work of others. Indeed the sometimes surprising interrelationships and analogies between different branches of mathematics are both the basis for many of its applications and the stimulus for further development. Secondly, different branches of mathematics appeal in different ways and require different talents. It is unlikely that all students at one university will have the same interests and aptitudes as their lecturers. Rather, they will only discover what their own interests and aptitudes are by being exposed to a broader range. Thirdly, many students of mathematics will become, not professional mathematicians, but scientists, engineers or schoolteachers. It is useful for them to have a clear understanding of the nature and extent of mathematics, and it is in the interests of mathematicians that there should be a body of people in the community who have this understanding.

The present book attempts to provide such an understanding of the nature and extent of mathematics. The connecting theme is the theory of numbers, at first sight one of the most abstruse and irrelevant branches of mathematics. Yet by exploring its many connections with other branches, we may obtain a broad picture. The topics chosen are not trivial and demand some effort on the part of the reader. As Euclid already said, there is no royal road. In general I have concentrated attention on those hard-won results which illuminate a wide area. If I am accused of picking the eyes out of some subjects, I have no defence except to say "But what beautiful eyes!"

The book is divided into two parts. Part A, which deals with elementary number theory, should be accessible to a first-year undergraduate. To provide a foundation for subsequent work, Chapter I contains the definitions and basic properties of various mathematical structures. However, the reader may simply skim through this chapter

and refer back to it later as required. Chapter V, on Hadamard's determinant problem, shows that elementary number theory may have unexpected applications.

Part B, which is more advanced, is intended to provide an undergraduate with some idea of the scope of mathematics today. The chapters in this part are largely independent, except that Chapter X depends on Chapter IX and Chapter XIII on Chapter XII.

Although much of the content of the book is common to any introductory work on number theory, I wish to draw attention to the discussion here of quadratic fields and elliptic curves. These are quite special cases of algebraic number fields and algebraic curves, and it may be asked why one should restrict attention to these special cases when the general cases are now well understood and may even be developed in parallel. My answers are as follows. First, to treat the general cases in full rigour requires a commitment of time which many will be unable to afford. Secondly, these special cases are those most commonly encountered and more constructive methods are available for them than for the general cases. There is yet another reason. Sometimes in mathematics a generalization is so simple and far-reaching that the special case is more fully understood as an instance of the generalization. For the topics mentioned, however, the generalization is more complex and is, in my view, more fully understood as a development from the special case.

At the end of each chapter of the book I have added a list of selected references, which will enable readers to travel further in their own chosen directions. Since the literature is voluminous, any such selection must be somewhat arbitrary, but I hope that mine may be found interesting and useful.

The computer revolution has made possible calculations on a scale and with a speed undreamt of a century ago. One consequence has been a considerable increase in 'experimental mathematics'—the search for patterns. This book, on the other hand, is devoted to 'theoretical mathematics'—the explanation of patterns. I do not wish to conceal the fact that the former usually precedes the latter. Nor do I wish to conceal the fact that some of the results here have been proved by the greatest minds of the past only after years of labour, and that their proofs have later been improved and simplified by many other mathematicians. Once obtained, however, a good proof organizes and provides understanding for a mass of computational data. Often it also suggests further developments.

The present book may indeed be viewed as a 'treasury of proofs'. We concentrate attention on this aspect of mathematics, not only because it is a distinctive feature of the subject, but also because we consider its exposition is better suited to a book than to a blackboard or a computer screen. In keeping with this approach, the proofs themselves have been chosen with some care and I hope that a few may be of interest even to those who are no longer students. Proofs which depend on general principles have been given preference over proofs which offer no particular insight.

Mathematics is a part of civilization and an achievement in which human beings may take some pride. It is not the possession of any one national, political or religious group and any attempt to make it so is ultimately destructive. At the present time there are strong pressures to make academic studies more 'relevant'. At the same time, however, staff at some universities are assessed by 'citation counts' and people are paid for giving lectures on chaos, for example, that are demonstrably rubbish.

The theory of numbers provides ample evidence that topics pursued for their own intrinsic interest can later find significant applications. I do not contend that curiosity has been the only driving force. More mundane motives, such as ambition or the necessity of earning a living, have also played a role. It is also true that mathematics pursued for the sake of applications has been of benefit to subjects such as number theory; there is a two-way trade. However, it shows a dangerous ignorance of history and of human nature to promote utility at the expense of spirit.

This book has its origin in a course of lectures which I gave at the Victoria University of Wellington, New Zealand, in 1975. The demands of my own research have hitherto prevented me from completing it, although I have continued to collect material. If it succeeds at all in conveying some idea of the power and beauty of mathematics, the labour of writing it will have been well worthwhile.

As with a previous book, I have to thank Helge Tverberg, who has read most of the manuscript and made many useful suggestions.

The first Phalanger Press edition of this book appeared in 2002. A revised edition, which was reissued by Springer in 2006, contained a number of changes. I removed an error in the statement and proof of Proposition II.12 and filled a gap in the proof of Proposition III.12. The statements of the Weil conjectures in Chapter IX and of a result of Heath-Brown in Chapter X were modified, following comments by J.-P. Serre. I also corrected a few misprints, made many small expository changes and expanded the index.

In the present edition I have made some more expository changes and have added a few references at the end of some chapters to take account of recent developments. For more detailed information the Internet has the advantage over a book. The reader is referred to the American Mathematical Society's MathSciNet (www.ams.org/mathscinet) and to The Number Theory Web maintained by Keith Matthews (www.maths.uq.edu.au/~krm/).

I am grateful to Springer for undertaking the commercial publication of my book and hope you will be also. Many of those who have contributed to the production of this new softcover edition are unknown to me, but among those who are I wish to thank especially Alicia de los Reyes and my sons Nicholas and Philip.

W.A. Coppel
May, 2009
Canberra, Australia

I

The Expanding Universe of Numbers

For many people, numbers must seem to be the essence of mathematics. *Number theory*, which is the subject of this book, is primarily concerned with the properties of one particular type of number, the 'whole numbers' or *integers*. However, there are many other types, such as complex numbers and p-adic numbers. Somewhat surprisingly, a knowledge of these other types turns out to be necessary for any deeper understanding of the integers.

In this introductory chapter we describe several such types (but defer the study of p-adic numbers to Chapter VI). *To embark on number theory proper the reader may proceed to Chapter II now* and refer back to the present chapter, via the Index, only as occasion demands.

When one studies the properties of various types of number, one becomes aware of formal similarities between different types. Instead of repeating the derivations of properties for each individual case, it is more economical – and sometimes actually clearer – to study their common algebraic structure. This algebraic structure may be shared by objects which one would not even consider as numbers.

There is a pedagogic difficulty here. Usually a property is discovered in one context and only later is it realized that it has wider validity. It may be more digestible to prove a result in the context of number theory and then simply point out its wider range of validity. Since this is a book on number theory, and many properties were first discovered in this context, we feel free to adopt this approach. However, to make the statements of such generalizations intelligible, in the latter part of this chapter we describe several basic algebraic structures. We do not attempt to study these structures in depth, but restrict attention to the simplest properties which throw light on the work of later chapters.

0 Sets, Relations and Mappings

The label '0' given to this section may be interpreted to stand for '0ptional'. We collect here some definitions of a logical nature which have become part of the common language of mathematics. Those who are not already familiar with this language, and who are repelled by its abstraction, should consult this section only when the need arises.

W.A. Coppel, *Number Theory: An Introduction to Mathematics*, Universitext,
DOI: 10.1007/978-0-387-89486-7_1, © Springer Science + Business Media, LLC 2009

We will not formally define a *set*, but will simply say that it is a collection of objects, which are called its *elements*. We write $a \in A$ if a is an element of the set A and $a \notin A$ if it is not.

A set may be specified by listing its elements. For example, $A = \{a, b, c\}$ is the set whose elements are a, b, c. A set may also be specified by characterizing its elements. For example,

$$A = \{x \in \mathbb{R} : x^2 < 2\}$$

is the set of all real numbers x such that $x^2 < 2$.

If two sets A, B have precisely the same elements, we say that they are *equal* and write $A = B$. (If A and B are not equal, we write $A \neq B$.) For example,

$$\{x \in \mathbb{R} : x^2 = 1\} = \{1, -1\}.$$

Just as it is convenient to admit 0 as a number, so it is convenient to admit the *empty set* $\emptyset$, which has no elements, as a set.

If every element of a set A is also an element of a set B we say that A is a *subset* of B, or that A is *included* in B, or that B *contains* A, and we write $A \subseteq B$. We say that A is a *proper* subset of B, and write $A \subset B$, if $A \subseteq B$ and $A \neq B$.

Thus $\emptyset \subseteq A$ for every set A and $\emptyset \subset A$ if $A \neq \emptyset$. Set inclusion has the following obvious properties:

(i) $A \subseteq A$;
(ii) *if* $A \subseteq B$ *and* $B \subseteq A$, *then* $A = B$;
(iii) *if* $A \subseteq B$ *and* $B \subseteq C$, *then* $A \subseteq C$.

For any sets A, B, the set whose elements are the elements of A or B (or both) is called the *union* or 'join' of A and B and is denoted by $A \cup B$:

$$A \cup B = \{x : x \in A \text{ or } x \in B\}.$$

The set whose elements are the common elements of A and B is called the *intersection* or 'meet' of A and B and is denoted by $A \cap B$:

$$A \cap B = \{x : x \in A \text{ and } x \in B\}.$$

If $A \cap B = \emptyset$, the sets A and B are said to be *disjoint*.

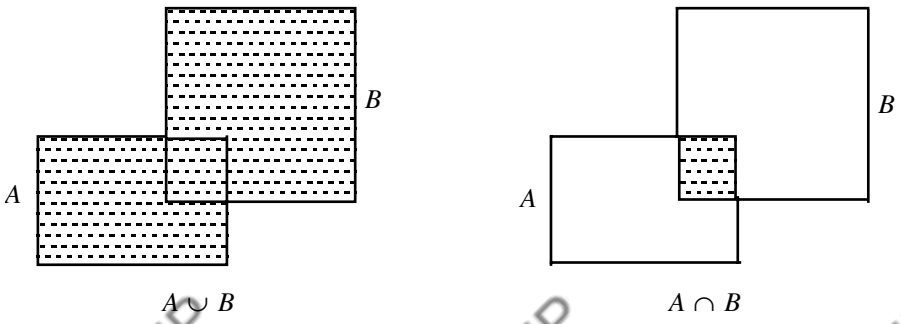


Fig. 1. Union and Intersection.

It is easily seen that union and intersection have the following algebraic properties:

$$\begin{aligned} A \cup A = A, \quad & A \cap A = A, \\ A \cup B = B \cup A, \quad & A \cap B = B \cap A, \\ (A \cup B) \cup C = A \cup (B \cup C), \quad & (A \cap B) \cap C = A \cap (B \cap C), \\ (A \cup B) \cap C = (A \cap C) \cup (B \cap C), \quad & (A \cap B) \cup C = (A \cup C) \cap (B \cup C). \end{aligned}$$

Set inclusion could have been defined in terms of either union or intersection, since $A \subseteq B$ is the same as $A \cup B = B$ and also the same as $A \cap B = A$.

For any sets A, B, the set of all elements of B which are not also elements of A is called the *difference* of B from A and is denoted by $B \backslash A$:

$$B \backslash A = \{x : x \in B \text{ and } x \notin A\}.$$

It is easily seen that

$$\begin{aligned} C \backslash (A \cup B) &= (C \backslash A) \cap (C \backslash B), \\ C \backslash (A \cap B) &= (C \backslash A) \cup (C \backslash B). \end{aligned}$$

An important special case is where all sets under consideration are subsets of a given universal set X. For any $A \subseteq X$, we have

$$\begin{aligned} \emptyset \cup A = A, \quad & \emptyset \cap A = \emptyset, \\ X \cup A = X, \quad & X \cap A = A. \end{aligned}$$

The set $X \backslash A$ is said to be the *complement* of A (in X) and may be denoted by A^{c} for fixed X. Evidently

$$\begin{aligned} \emptyset^{\mathrm{c}} = X, \quad & X^{\mathrm{c}} = \emptyset, \\ A \cup A^{\mathrm{c}} = X, \quad & A \cap A^{\mathrm{c}} = \emptyset, \\ & (A^{\mathrm{c}})^{\mathrm{c}} = A. \end{aligned}$$

By taking $C = X$ in the previous relations for differences, we obtain 'De Morgan's laws':

$$(A \cup B)^{\mathrm{c}} = A^{\mathrm{c}} \cap B^{\mathrm{c}}, (A \cap B)^{\mathrm{c}} = A^{\mathrm{c}} \cup B^{\mathrm{c}}.$$

Since $A \cap B = (A^{\mathrm{c}} \cup B^{\mathrm{c}})^{\mathrm{c}}$, set intersection can be defined in terms of unions and complements. Alternatively, since $A \cup B = (A^{\mathrm{c}} \cap B^{\mathrm{c}})^{\mathrm{c}}$, set union can be defined in terms of intersections and complements.

For any sets A, B, the set of all ordered pairs (a, b) with $a \in A$ and $b \in B$ is called the (*Cartesian*) *product* of A by B and is denoted by $A \times B$.

Similarly one can define the product of more than two sets. We mention only one special case. For any positive integer n, we write A^n instead of $A \times \cdots \times A$ for the set of all (ordered) n-*tuples* $(a_1, \ldots, a_n)$ with $a_j \in A$ $(1 \leq j \leq n)$. We call a_j the j-th *coordinate* of the n-tuple.

A *binary relation* on a set A is just a subset R of the product set $A \times A$. For any $a, b \in A$, we write aRb if $(a, b) \in R$. A binary relation R on a set A is said to be

reflexive if aRa for every $a \in A$;
symmetric if bRa whenever aRb;
transitive if aRc whenever aRb and bRc.

It is said to be an *equivalence relation* if it is reflexive, symmetric and transitive.

If R is an equivalence relation on a set A and $a \in A$, the *equivalence class* R_a of a is the set of all $x \in A$ such that xRa. Since R is reflexive, $a \in R_a$. Since R is symmetric, $b \in R_a$ implies $a \in R_b$. Since R is transitive, $b \in R_a$ implies $R_b \subseteq R_a$. It follows that, for all $a, b \in A$, either $R_a = R_b$ or $R_a \cap R_b = \emptyset$.

A *partition* $\mathscr{C}$ of a set A is a collection of nonempty subsets of A such that each element of A is an element of exactly one of the subsets in $\mathscr{C}$.

Thus the distinct equivalence classes corresponding to a given equivalence relation on a set A form a partition of A. It is not difficult to see that, conversely, if $\mathscr{C}$ is a partition of A, then an equivalence relation R is defined on A by taking R to be the set of all $(a, b) \in A \times A$ for which a and b are elements of the same subset in the collection $\mathscr{C}$.

Let A and B be nonempty sets. A *mapping* f of A into B is a subset of $A \times B$ with the property that, for each $a \in A$, there is a unique $b \in B$ such that $(a, b) \in f$. We write $f(a) = b$ if $(a, b) \in f$, and say that b is the *image* of a under f or that b is the *value* of f at a. We express that f is a mapping of A into B by writing $f: A \to B$ and we put

$$f(A) = \{f(a): a \in A\}.$$

The term *function* is often used instead of 'mapping', especially when A and B are sets of real or complex numbers, and 'mapping' itself is often abbreviated to *map*.

If f is a mapping of A into B, and if A' is a nonempty subset of A, then the *restriction* of f to A' is the set of all $(a, b) \in f$ with $a \in A'$.

The *identity map* i_A of a nonempty set A into itself is the set of all ordered pairs (a, a) with $a \in A$.

If f is a mapping of A into B, and g a mapping of B into C, then the *composite mapping* $g \circ f$ of A into C is the set of all ordered pairs (a, c), where $c = g(b)$ and $b = f(a)$. Composition of mappings is associative, i.e. if h is a mapping of C into D, then

$$(h \circ g) \circ f = h \circ (g \circ f).$$

The identity map has the obvious properties $f \circ i_A = f$ and $i_B \circ f = f$.

Let A, B be nonempty sets and $f: A \to B$ a mapping of A into B. The mapping f is said to be 'one-to-one' or *injective* if, for each $b \in B$, there exists at most one $a \in A$ such that $(a, b) \in f$. The mapping f is said to be 'onto' or *surjective* if, for each $b \in B$, there exists at least one $a \in A$ such that $(a, b) \in f$. If f is both injective and surjective, then it is said to be *bijective* or a 'one-to-one correspondence'. The nouns *injection, surjection* and *bijection* are also used instead of the corresponding adjectives.

It is not difficult to see that f is injective if and only if there exists a mapping $g: B \to A$ such that $g \circ f = i_A$, and surjective if and only if there exists a mapping $h: B \to A$ such that $f \circ h = i_B$. Furthermore, if f is bijective, then g and h are

unique and equal. Thus, for any bijective map $f: A \to B$, there is a unique *inverse* map $f^{-1}: B \to A$ such that $f^{-1} \circ f = i_A$ and $f \circ f^{-1} = i_B$.

If $f: A \to B$ and $g: B \to C$ are both bijective maps, then $g \circ f: A \to C$ is also bijective and

$$(g \circ f)^{-1} = f^{-1} \circ g^{-1}.$$

1 Natural Numbers

The natural numbers are the numbers usually denoted by $1, 2, 3, 4, 5, \ldots$. However, other notations are also used, e.g. for the chapters of this book. Although one notation may have considerable practical advantages over another, it is the properties of the natural numbers which are basic.

The following system of axioms for the natural numbers was essentially given by Dedekind (1888), although it is usually attributed to Peano (1889):

The natural numbers are the elements of a set $\mathbb{N}$*, with a distinguished element* 1 (*one*) *and map* $S: \mathbb{N} \to \mathbb{N}$*, such that*

(N1) *S is injective*, i.e. *if* $m, n \in \mathbb{N}$ *and* $m \neq n$*, then* $S(m) \neq S(n)$;
(N2) $1 \notin S(\mathbb{N})$;
(N3) *if* $M \subseteq \mathbb{N}$, $1 \in M$ *and* $S(M) \subseteq M$*, then* $M = \mathbb{N}$.

The element $S(n)$ of $\mathbb{N}$ is called the *successor* of n. The axioms are satisfied by $\{1, 2, 3, \ldots\}$ if we take $S(n)$ to be the element immediately following the element n.

It follows readily from the axioms that 1 is the only element of $\mathbb{N}$ which is not in $S(\mathbb{N})$. For, if $M = S(\mathbb{N}) \cup \{1\}$, then $M \subseteq \mathbb{N}$, $1 \in M$ and $S(M) \subseteq M$. Hence, by **(N3)**, $M = \mathbb{N}$.

It also follows from the axioms that $S(n) \neq n$ for every $n \in \mathbb{N}$. For let M be the set of all $n \in \mathbb{N}$ such that $S(n) \neq n$. By **(N2)**, $1 \in M$. If $n \in M$ and $n' = S(n)$ then, by **(N1)**, $S(n') \neq n'$. Thus $S(M) \subseteq M$ and hence, by **(N3)**, $M = \mathbb{N}$.

The axioms **(N1)–(N3)** actually determine $\mathbb{N}$ up to 'isomorphism'. We will deduce this as a corollary of the following general *recursion theorem*:

Proposition 1 *Given a set A, an element* a_1 *of A and a map* $T: A \to A$*, there exists exactly one map* $\varphi: \mathbb{N} \to A$ *such that* $\varphi(1) = a_1$ *and*

$$\varphi(S(n)) = T\varphi(n) \quad \textit{for every } n \in \mathbb{N}.$$

Proof We show first that there is at most one map with the required properties. Let φ_1 and φ_2 be two such maps, and let M be the set of all $n \in \mathbb{N}$ such that

$$\varphi_1(n) = \varphi_2(n).$$

Evidently $1 \in M$. If $n \in M$, then also $S(n) \in M$, since

$$\varphi_1(S(n)) = T\varphi_1(n) = T\varphi_2(n) = \varphi_2(S(n)).$$

Hence, by **(N3)**, $M = \mathbb{N}$. That is, $\varphi_1 = \varphi_2$.

We now show that there exists such a map φ. Let $\mathscr{C}$ be the collection of all subsets C of $\mathbb{N} \times A$ such that $(1, a_1) \in C$ and such that if $(n, a) \in C$, then also $(S(n), T(a)) \in C$. The collection $\mathscr{C}$ is not empty, since it contains $\mathbb{N} \times A$. Moreover, since every set in $\mathscr{C}$ contains $(1, a_1)$, the intersection D of all sets $C \in \mathscr{C}$ is not empty. It is easily seen that actually $D \in \mathscr{C}$. By its definition, however, no proper subset of D is in $\mathscr{C}$.

Let M be the set of all $n \in \mathbb{N}$ such that $(n, a) \in D$ for exactly one $a \in A$ and, for any $n \in M$, define $\varphi(n)$ to be the unique $a \in A$ such that $(n, a) \in D$. If $M = \mathbb{N}$, then $\varphi(1) = a_1$ and $\varphi(S(n)) = T\varphi(n)$ for all $n \in \mathbb{N}$. Thus we need only show that $M = \mathbb{N}$. As usual, we do this by showing that $1 \in M$ and that $n \in M$ implies $S(n) \in M$.

We have $(1, a_1) \in D$. Assume $(1, a') \in D$ for some $a' \neq a_1$. If $D' = D \backslash \{(1, a')\}$, then $(1, a_1) \in D'$. Moreover, if $(n, a) \in D'$ then $(S(n), T(a)) \in D'$, since $(S(n), T(a)) \in D$ and $(S(n), T(a)) \neq (1, a')$. Hence $D' \in \mathscr{C}$. But this is a contradiction, since D' is a proper subset of D. We conclude that $1 \in M$.

Suppose now that $n \in M$ and let a be the unique element of A such that $(n, a) \in D$. Then $(S(n), T(a)) \in D$, since $D \in \mathscr{C}$. Assume that $(S(n), a'') \in D$ for some $a'' \neq T(a)$ and put $D'' = D \backslash \{(S(n), a'')\}$. Then $(S(n), T(a)) \in D''$ and $(1, a_1) \in D''$. For any $(m, b) \in D''$ we have $(S(m), T(b)) \in D$. If $(S(m), T(b)) = (S(n), a'')$, then $S(m) = S(n)$ and $T(b) = a'' \neq T(a)$, which implies $m = n$ and $b \neq a$. Thus D contains both (n, b) and (n, a), which contradicts $n \in M$. Hence $(S(m), T(b)) \neq (S(n), a'')$, and so $(S(m), T(b)) \in D''$. But then $D'' \in \mathscr{C}$, which is also a contradiction, since D'' is a proper subset of D. We conclude that $S(n) \in M$. □

Corollary 2 *If the axioms* **(N1)–(N3)** *are also satisfied by a set* $\mathbb{N}'$ *wth element* $1'$ *and map* $S' : \mathbb{N}' \to \mathbb{N}'$, *then there exists a bijective map* φ *of* $\mathbb{N}$ *onto* $\mathbb{N}'$ *such that* $\varphi(1) = 1'$ *and*

$$\varphi(S(n)) = S'\varphi(n) \quad \textit{for every } n \in \mathbb{N}.$$

Proof By taking $A = \mathbb{N}'$, $a_1 = 1'$ and $T = S'$ in Proposition 1, we see that there exists a unique map $\varphi : \mathbb{N} \to \mathbb{N}'$ such that $\varphi(1) = 1'$ and

$$\varphi(S(n)) = S'\varphi(n) \quad \text{for every } n \in \mathbb{N}.$$

By interchanging $\mathbb{N}$ and $\mathbb{N}'$, we see also that there exists a unique map $\psi : \mathbb{N}' \to \mathbb{N}$ such that $\psi(1') = 1$ and

$$\psi(S'(n')) = S\psi(n') \quad \text{for every } n' \in \mathbb{N}'.$$

The composite map $\chi = \psi \circ \varphi$ of $\mathbb{N}$ into $\mathbb{N}$ has the properties $\chi(1) = 1$ and $\chi(S(n)) = S\chi(n)$ for every $n \in \mathbb{N}$. But, by Proposition 1 again, χ is uniquely determined by these properties. Hence $\psi \circ \varphi$ is the identity map on $\mathbb{N}$, and similarly $\varphi \circ \psi$ is the identity map on $\mathbb{N}'$. Consequently φ is a bijection. □

We can also use Proposition 1 to define addition and multiplication of natural numbers. By Proposition 1, for each $m \in \mathbb{N}$ there exists a unique map $s_m : \mathbb{N} \to \mathbb{N}$ such that

$$s_m(1) = S(m), \quad s_m(S(n)) = Ss_m(n) \quad \text{for every } n \in \mathbb{N}.$$

We define the *sum* of m and n to be

$$m + n = s_m(n).$$

It is not difficult to deduce from this definition and the axioms **(N1)–(N3)** the usual rules for *addition*: *for all* $a, b, c \in \mathbb{N}$,

(A1) *if* $a + c = b + c$, *then* $a = b$; (cancellation law)
(A2) $a + b = b + a$; (commutative law)
(A3) $(a + b) + c = a + (b + c)$. (associative law)

By way of example, we prove the cancellation law. Let M be the set of all $c \in \mathbb{N}$ such that $a + c = b + c$ only if $a = b$. Then $1 \in M$, since $s_a(1) = s_b(1)$ implies $S(a) = S(b)$ and hence $a = b$. Suppose $c \in M$. If $a + S(c) = b + S(c)$, i.e. $s_a(S(c)) = s_b(S(c))$, then $Ss_a(c) = Ss_b(c)$ and hence, by **(N1)**, $s_a(c) = s_b(c)$. Since $c \in M$, this implies $a = b$. Thus also $S(c) \in M$. Hence, by **(N3)**, $M = \mathbb{N}$.

We now show that

$$m + n \neq n \quad \text{for all } m, n \in \mathbb{N}.$$

For a given $m \in \mathbb{N}$, let M be the set of all $n \in \mathbb{N}$ such that $m + n \neq n$. Then $1 \in M$ since, by **(N2)**, $s_m(1) = S(m) \neq 1$. If $n \in M$, then $s_m(n) \neq n$ and hence, by **(N1)**,

$$s_m(S(n)) = Ss_m(n) \neq S(n).$$

Hence, by **(N3)**, $M = \mathbb{N}$.

By Proposition 1 again, for each $m \in \mathbb{N}$ there exists a unique map $p_m : \mathbb{N} \to \mathbb{N}$ such that

$$\begin{aligned} p_m(1) &= m, \\ p_m(S(n)) &= s_m(p_m(n)) \quad \text{for every } n \in \mathbb{N}. \end{aligned}$$

We define the *product* of m and n to be

$$m \cdot n = p_m(n).$$

From this definition and the axioms **(N1)–(N3)** we may similarly deduce the usual rules for *multiplication*: *for all* $a, b, c \in \mathbb{N}$,

(M1) *if* $a \cdot c = b \cdot c$, *then* $a = b$; (cancellation law)
(M2) $a \cdot b = b \cdot a$; (commutative law)
(M3) $(a \cdot b) \cdot c = a \cdot (b \cdot c)$; (associative law)
(M4) $a \cdot 1 = a$. (identity element)

Furthermore, addition and multiplication are connected by

(AM1) $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$. (distributive law)

As customary, we will often omit the dot when writing products and we will give multiplication precedence over addition. With these conventions the distributive law becomes simply

$$a(b + c) = ab + ac.$$

We show next how a relation of order may be defined on the set $\mathbb{N}$. For any $m, n \in \mathbb{N}$, we say that m *is less than* n, and write $m < n$, if

$$m + m' = n \quad \text{for some } m' \in \mathbb{N}.$$

Evidently $m < S(m)$ for every $m \in \mathbb{N}$, since $S(m) = m + 1$. Also, if $m < n$, then either $S(m) = n$ or $S(m) < n$. For suppose $m + m' = n$. If $m' = 1$, then $S(m) = n$. If $m' \neq 1$, then $m' = m'' + 1$ for some $m'' \in \mathbb{N}$ and

$$S(m) + m'' = (m + 1) + m'' = m + (1 + m'') = m + m' = n.$$

Again, if $n \neq 1$, then $1 < n$, since the set consisting of 1 and all $n \in \mathbb{N}$ such that $1 < n$ contains 1 and contains $S(n)$ if it contains n.

It will now be shown that the relation '$<$' induces a *total order* on $\mathbb{N}$, which is compatible with both addition and multiplication: *for all* $a, b, c \in \mathbb{N}$,

(O1) *if* $a < b$ *and* $b < c$, *then* $a < c$; (transitive law)

(O2) *one and only one of the following alternatives holds*:

$$a < b, a = b, b < a; \quad \text{(law of trichotomy)}$$

(O3) $a + c < b + c$ *if and only if* $a < b$;

(O4) $ac < bc$ *if and only if* $a < b$.

The relation **(O1)** follows directly from the associative law for addition. We now prove **(O2)**. If $a < b$ then, for some $a' \in \mathbb{N}$,

$$b = a + a' = a' + a \neq a.$$

Together with **(O1)**, this shows that at most one of the three alternatives in **(O2)** holds.

For a given $a \in \mathbb{N}$, let M be the set of all $b \in \mathbb{N}$ such that at least one of the three alternatives in **(O2)** holds. Then $1 \in M$, since $1 < a$ if $a \neq 1$. Suppose now that $b \in M$. If $a = b$, then $a < S(b)$. If $a < b$, then again $a < S(b)$, by **(O1)**. If $b < a$, then either $S(b) = a$ or $S(b) < a$. Hence also $S(b) \in M$. Consequently, by **(N3)**, $M = \mathbb{N}$. This completes the proof of **(O2)**.

It follows from the associative and commutative laws for addition that, if $a < b$, then $a + c < b + c$. On the other hand, by using also the cancellation law we see that if $a + c < b + c$, then $a < b$.

It follows from the distributive law that, if $a < b$, then $ac < bc$. Finally, suppose $ac < bc$. Then $a \neq b$ and hence, by **(O2)**, either $a < b$ or $b < a$. Since $b < a$ would imply $bc < ac$, by what we have just proved, we must actually have $a < b$.

The law of trichotomy **(O2)** implies that, for given $m, n \in \mathbb{N}$, the equation

$$m + x = n$$

has a solution $x \in \mathbb{N}$ only if $m < n$.

As customary, we write $a \leq b$ to denote either $a < b$ or $a = b$. Also, it is sometimes convenient to write $b > a$ instead of $a < b$, and $b \geq a$ instead of $a \leq b$.

A subset M of $\mathbb{N}$ is said to have a *least element* m' if $m' \in M$ and $m' \leq m$ for every $m \in M$. The least element m' is uniquely determined, if it exists, by **(O2)**. By what we have already proved, 1 is the least element of $\mathbb{N}$.

Proposition 3 *Any nonempty subset M of $\mathbb{N}$ has a least element.*

Proof Assume that some nonempty subset M of $\mathbb{N}$ does not have a least element. Then $1 \notin M$, since 1 is the least element of $\mathbb{N}$. Let L be the set of all $l \in \mathbb{N}$ such that $l < m$ for every $m \in M$. Then L and M are disjoint and $1 \in L$. If $l \in L$, then $S(l) \leq m$ for every $m \in M$. Since M does not have a least element, it follows that $S(l) \notin M$. Thus $S(l) < m$ for every $m \in M$, and so $S(l) \in L$. Hence, by **(N3)**, $L = \mathbb{N}$. Since $L \cap M = \emptyset$, this is a contradiction. □

The method of *proof by induction* is a direct consequence of the axioms defining $\mathbb{N}$. Suppose that with each $n \in \mathbb{N}$ there is associated a proposition P_n. To show that P_n is true for every $n \in \mathbb{N}$, we need only show that P_1 is true and that P_{n+1} is true if P_n is true.

Proposition 3 provides an alternative approach. To show that P_n is true for every $n \in \mathbb{N}$, we need only show that if P_m is false for some m, then P_l is false for some $l < m$. For then the set of all $n \in \mathbb{N}$ for which P_n is false has no least element and consequently is empty.

For any $n \in \mathbb{N}$, we denote by I_n the set of all $m \in \mathbb{N}$ such that $m \leq n$. Thus $I_1 = \{1\}$ and $S(n) \notin I_n$. It is easily seen that

$$I_{S(n)} = I_n \cup \{S(n)\}.$$

Also, for any $p \in I_{S(n)}$, there exists a bijective map f_p of I_n onto $I_{S(n)} \backslash \{p\}$. For, if $p = S(n)$ we can take f_p to be the identity map on I_n, and if $p \in I_n$ we can take f_p to be the map defined by

$$f_p(p) = S(n), \; f_p(m) = m \quad \text{if } m \in I_n \backslash \{p\}.$$

Proposition 4 *For any $m, n \in \mathbb{N}$, if a map $f : I_m \to I_n$ is injective and $f(I_m) \neq I_n$, then $m < n$.*

Proof The result certainly holds when $m = 1$, since $I_1 = \{1\}$. Let M be the set of all $m \in \mathbb{N}$ for which the result holds. We need only show that if $m \in M$, then also $S(m) \in M$.

Let $f : I_{S(m)} \to I_n$ be an injective map such that $f(I_{S(m)}) \neq I_n$ and choose $p \in I_n \backslash f(I_{S(m)})$. The restriction g of f to I_m is also injective and $g(I_m) \neq I_n$. Since $m \in M$, it follows that $m < n$. Assume $S(m) = n$. Then there exists a bijective map g_p of $I_{S(m)} \backslash \{p\}$ onto I_m. The composite map $h = g_p \circ f$ maps $I_{S(m)}$ into I_m and is injective. Since $m \in M$, we must have $h(I_m) = I_m$. But, since $h(S(m)) \in I_m$ and h is injective, this is a contradiction. Hence $S(m) < n$ and, since this holds for every f, $S(m) \in M$. □

Proposition 5 *For any $m, n \in \mathbb{N}$, if a map $f : I_m \to I_n$ is not injective and $f(I_m) = I_n$, then $m > n$.*

Proof The result holds vacuously when $m = 1$, since any map $f : I_1 \to I_n$ is injective. Let M be the set of all $m \in \mathbb{N}$ for which the result holds. We need only show that if $m \in M$, then also $S(m) \in M$.

Let $f: I_{S(m)} \to I_n$ be a map such that $f(I_{S(m)}) = I_n$ which is not injective. Then there exist $p, q \in I_{S(m)}$ with $p \neq q$ and $f(p) = f(q)$. We may choose the notation so that $q \in I_m$. If f_p is a bijective map of I_m onto $I_{S(m)} \backslash \{p\}$, then the composite map $h = f \circ f_p$ maps I_m onto I_n. If it is not injective then $m > n$, since $m \in M$, and hence also $S(m) > n$. If h is injective, then it is bijective and has a bijective inverse $h^{-1}: I_n \to I_m$. Since $h^{-1}(I_n)$ is a proper subset of $I_{S(m)}$, it follows from Proposition 4 that $n < S(m)$. Hence $S(m) \in M$. □

Propositions 4 and 5 immediately imply

Corollary 6 *For any $n \in \mathbb{N}$, a map $f: I_n \to I_n$ is injective if and only if it is surjective.*

Corollary 7 *If a map $f: I_m \to I_n$ is bijective, then $m = n$.*

Proof By Proposition 4, $m < S(n)$, i.e. $m \leq n$. Replacing f by f^{-1}, we obtain in the same way $n \leq m$. Hence $m = n$. □

A set E is said to be *finite* if there exists a bijective map $f: E \to I_n$ for some $n \in \mathbb{N}$. Then n is uniquely determined, by Corollary 7. We call it the *cardinality* of E and denote it by $\#(E)$.

It is readily shown that if E is a finite set and F a proper subset of E, then F is also finite and $\#(F) < \#(E)$. Again, if E and F are disjoint finite sets, then their union $E \cup F$ is also finite and $\#(E \cup F) = \#(E) + \#(F)$. Furthermore, for any finite sets E and F, the product set $E \times F$ is also finite and $\#(E \times F) = \#(E) \cdot \#(F)$.

Corollary 6 implies that, for any finite set E, a map $f: E \to E$ is injective if and only if it is surjective. This is a precise statement of the so-called *pigeonhole principle*.

A set E is said to be *countably infinite* if there exists a bijective map $f: E \to \mathbb{N}$. Any countably infinite set may be bijectively mapped onto a proper subset F, since $\mathbb{N}$ is bijectively mapped onto a proper subset by the successor map S. Thus a map $f: E \to E$ of an infinite set E may be injective, but not surjective. It may also be surjective, but not injective; an example is the map $f: \mathbb{N} \to \mathbb{N}$ defined by $f(1) = 1$ and, for $n \neq 1$, $f(n) = m$ if $S(m) = n$.

2 Integers and Rational Numbers

The concept of number will now be extended. The natural numbers $1, 2, 3, \ldots$ suffice for counting purposes, but for bank balance purposes we require the larger set $\ldots, -2, -1, 0, 1, 2, \ldots$ of integers. (From this point of view, -2 is not so 'unnatural'.) An important reason for extending the concept of number is the greater freedom it gives us. In the realm of natural numbers the equation $a + x = b$ has a solution if and only if $b > a$; in the extended realm of integers it will always have a solution.

Rather than introduce a new set of axioms for the integers, we will define them in terms of natural numbers. Intuitively, an integer is the difference $m - n$ of two natural numbers m, n, with addition and multiplication defined by

$$(m - n) + (p - q) = (m + p) - (n + q),$$
$$(m - n) \cdot (p - q) = (mp + nq) - (mq + np).$$

However, two other natural numbers m', n' may have the same difference as m, n, and anyway what does $m - n$ mean if $m < n$? To make things precise, we proceed in the following way.

Consider the set $\mathbb{N} \times \mathbb{N}$ of all ordered pairs of natural numbers. For any two such ordered pairs, (m, n) and (m', n'), we write

$$(m, n) \sim (m', n') \quad \text{if } m + n' = m' + n.$$

We will show that this is an *equivalence relation*. It follows at once from the definition that $(m, n) \sim (m, n)$ (reflexive law) and that $(m, n) \sim (m', n')$ implies $(m', n') \sim (m, n)$ (symmetric law). It remains to prove the transitive law:

$$(m, n) \sim (m', n') \text{ and } (m', n') \sim (m'', n'') \text{ imply } (m, n) \sim (m'', n'').$$

This follows from the commutative, associative and cancellation laws for addition in $\mathbb{N}$. For we have

$$m + n' = m' + n, \quad m' + n'' = m'' + n',$$

and hence

$$(m + n') + n'' = (m' + n) + n'' = (m' + n'') + n = (m'' + n') + n.$$

Thus

$$(m + n'') + n' = (m'' + n) + n',$$

and so $m + n'' = m'' + n$.

The equivalence class containing $(1, 1)$ evidently consists of all pairs (m, n) with $m = n$.

We define an *integer* to be an equivalence class of ordered pairs of natural numbers and, as is now customary, we denote the set of all integers by $\mathbb{Z}$.

Addition of integers is defined componentwise:

$$(m, n) + (p, q) = (m + p, n + q).$$

To justify this definition we must show that it does not depend on the choice of representatives within an equivalence class, i.e. that

$$(m, n) \sim (m', n') \text{ and } (p, q) \sim (p', q') \text{ imply } (m + p, n + q) \sim (m' + p', n' + q').$$

However, if

$$m + n' = m' + n, \quad p + q' = p' + q,$$

then

$$\begin{aligned}(m + p) + (n' + q') &= (m + n') + (p + q') \\ &= (m' + n) + (p' + q) = (m' + p') + (n + q).\end{aligned}$$

It follows at once from the corresponding properties of natural numbers that, also in $\mathbb{Z}$, addition satisfies the commutative law **(A2)** and the associative law **(A3)**. Moreover, the equivalence class 0 (*zero*) containing (1,1) is an *identity element* for addition:

(A4) $a + 0 = a$ *for every* a.

Furthermore, the equivalence class containing (n, m) is an *additive inverse* for the equivalence containing (m, n):

(A5) *for each a, there exists* $-a$ *such that* $a + (-a) = 0$.

From these properties we can now obtain

Proposition 8 *For all* $a, b \in \mathbb{Z}$, *the equation* $a + x = b$ *has a unique solution* $x \in \mathbb{Z}$.

Proof It is clear that $x = (-a) + b$ is a solution. Moreover, this solution is unique, since if $a + x = a + x'$ then, by adding $-a$ to both sides, we obtain $x = x'$. □

Proposition 8 shows that the cancellation law **(A1)** is a consequence of **(A2)–(A5)**. It also immediately implies

Corollary 9 *For each* $a \in \mathbb{Z}$, 0 *is the only element such that* $a+0 = a$, $-a$ *is uniquely determined by a, and* $a = -(-a)$.

As usual, we will henceforth write $b - a$ instead of $b + (-a)$.

Multiplication of integers is defined by

$$(m, n) \cdot (p, q) = (mp + nq, mq + np).$$

To justify this definition we must show that $(m, n) \sim (m', n')$ and $(p, q) \sim (p', q')$ imply

$$(mp + nq, mq + np) \sim (m'p' + n'q', m'q' + n'p').$$

From $m + n' = m' + n$, by multiplying by p and q we obtain

$$\begin{aligned} mp + n'p &= m'p + np, \\ m'q + nq &= mq + n'q, \end{aligned}$$

and from $p + q' = p' + q$, by multiplying by m' and n' we obtain

$$\begin{aligned} m'p + m'q' &= m'p' + m'q, \\ n'p' + n'q &= n'p + n'q'. \end{aligned}$$

Adding these four equations and cancelling the terms common to both sides, we get

$$(mp + nq) + (m'q' + n'p') = (m'p' + n'q') + (mq + np),$$

as required.

It is easily verified that, also in $\mathbb{Z}$, multiplication satisfies the commutative law **(M2)** and the associative law **(M3)**. Moreover, the distributive law **(AM1)** holds and, if 1 is the equivalence class containing $(1 + 1, 1)$, then **(M4)** also holds. (In practice it does not cause confusion to denote identity elements of $\mathbb{N}$ and $\mathbb{Z}$ by the same symbol.)

Proposition 10 *For every $a \in \mathbb{Z}$, $a \cdot 0 = 0$.*

Proof We have

$$a \cdot 0 = a \cdot (0 + 0) = a \cdot 0 + a \cdot 0.$$

Adding $-(a \cdot 0)$ to both sides, we obtain the result. □

Proposition 10 could also have been derived directly from the definitions, but we prefer to view it as a consequence of the properties which have been labelled.

Corollary 11 *For all $a, b \in \mathbb{Z}$,*

$$a(-b) = -(ab), (-a)(-b) = ab.$$

Proof The first relation follows from

$$ab + a(-b) = a \cdot 0 = 0,$$

and the second relation follows from the first, since $c = -(-c)$. □

By the definitions of 0 and 1 we also have

(AM2) $1 \neq 0$.

(In fact $1 = 0$ would imply $a = 0$ for every a, since $a \cdot 1 = a$ and $a \cdot 0 = 0$.)

We will say that an integer a is *positive* if it is represented by an ordered pair (m, n) with $n < m$. This definition does not depend on the choice of representative. For if $n < m$ and $m + n' = m' + n$, then $m + n' < m' + m$ and hence $n' < m'$.

We will denote by P the set of all positive integers. The law of trichotomy **(O2)** for natural numbers immediately implies

(P1) *for every a, one and only one of the following alternatives holds*:

$$a \in P, \quad a = 0, \quad -a \in P.$$

We say that an integer is *negative* if it has the form $-a$, where $a \in P$, and we denote by $-P$ the set of all negative integers. Since $a = -(-a)$, **(P1)** says that $\mathbb{Z}$ is the disjoint union of the sets P, $\{0\}$ and $-P$.

From the property **(O3)** of natural numbers we immediately obtain

(P2) *if $a \in P$ and $b \in P$, then $a + b \in P$.*

Furthermore, we have

(P3) *if $a \in P$ and $b \in P$, then $a \cdot b \in P$.*

To prove this we need only show that if m, n, p, q are natural numbers such that $n < m$ and $q < p$, then

$$mq + np < mp + nq.$$

Since $q < p$, there exists a natural number q' such that $q + q' = p$. But then $nq' < mq'$, since $n < m$, and hence

$$mq + np = (m + n)q + nq' < (m + n)q + mq' = mp + nq.$$

We may write **(P2)** and **(P3)** symbolically in the form

$$P + P \subseteq P, \quad P \cdot P \subseteq P.$$

We now show that there are no *divisors of zero* in $\mathbb{Z}$:

Proposition 12 *If $a \neq 0$ and $b \neq 0$, then $ab \neq 0$.*

Proof By **(P1)**, either a or $-a$ is positive, and either b or $-b$ is positive. If $a \in P$ and $b \in P$ then $ab \in P$, by **(P3)**, and hence $ab \neq 0$, by **(P1)**. If $a \in P$ and $-b \in P$, then $a(-b) \in P$. Hence $ab = -(a(-b)) \in -P$ and $ab \neq 0$. Similarly if $-a \in P$ and $b \in P$. Finally, if $-a \in P$ and $-b \in P$, then $ab = (-a)(-b) \in P$ and again $ab \neq 0$. □

The proof of Proposition 12 also shows that any nonzero square is positive:

Proposition 13 *If $a \neq 0$, then $a^2 := aa \in P$.*

It follows that $1 \in P$, since $1 \neq 0$ and $1^2 = 1$.

The set P of positive integers induces an order relation in $\mathbb{Z}$. Write

$$a < b \quad \text{if } b - a \in P,$$

so that $a \in P$ if and only if $0 < a$. From this definition and the properties of P it follows that the order properties **(O1)–(O3)** hold also in $\mathbb{Z}$, and that **(O4)** holds in the modified form:

(O4)′ *if $0 < c$, then $ac < bc$ if and only if $a < b$.*

We now show that we can represent any $a \in \mathbb{Z}$ in the form $a = b - c$, where $b, c \in P$. In fact, if $a = 0$, we can take $b = 1$ and $c = 1$; if $a \in P$, we can take $b = a + 1$ and $c = 1$; and if $-a \in P$, we can take $b = 1$ and $c = 1 - a$.

An element a of $\mathbb{Z}$ is said to be a *lower bound* for a subset X of $\mathbb{Z}$ if $a \leq x$ for every $x \in X$. Proposition 3 immediately implies that if a subset of $\mathbb{Z}$ has a lower bound, then it has a least element.

For any $n \in \mathbb{N}$, let n' be the integer represented by $(n + 1, 1)$. Then $n' \in P$. We are going to study the map $n \to n'$ of $\mathbb{N}$ into P. The map is injective, since $n' = m'$ implies $n = m$. It is also surjective, since if $a \in P$ is represented by (m, n), where $n < m$, then it is also represented by $(p + 1, 1)$, where $p \in \mathbb{N}$ satisfies $n + p = m$. It is easily verified that the map preserves sums and products:

$$(m + n)' = m' + n', \quad (mn)' = m'n'.$$

Since $1' = 1$, it follows that $S(n)' = n' + 1$. Furthermore, we have

$$m' < n' \quad \text{if and only if } m < n.$$

Thus the map $n \to n'$ establishes an 'isomorphism' of $\mathbb{N}$ with P. In other words, P is a copy of $\mathbb{N}$ situated within $\mathbb{Z}$. By identifying n with n', we may regard $\mathbb{N}$ itself as a subset of $\mathbb{Z}$ (and stop talking about P). Then 'natural number' is the same as 'positive integer' and any integer is the difference of two natural numbers.

Number theory, in its most basic form, is the study of the properties of the set $\mathbb{Z}$ of integers. It will be considered in some detail in later chapters of this book, but to relieve the abstraction of the preceding discussion we consider here the *division algorithm:*

Proposition 14 *For any integers a, b with $a > 0$, there exist unique integers q, r such that*

$$b = qa + r, \quad 0 \leq r < a.$$

Proof We consider first uniqueness. Suppose

$$qa + r = q'a + r', \quad 0 \leq r, r' < a.$$

If $r < r'$, then from

$$(q - q')a = r' - r,$$

we obtain first $q > q'$ and then $r' - r \geq a$, which is a contradiction. If $r' < r$, we obtain a contradiction similarly. Hence $r = r'$, which implies $q = q'$.

We consider next existence. Let S be the set of all integers $y \geq 0$ which can be represented in the form $y = b - xa$ for some $x \in \mathbb{Z}$. The set S is not empty, since it contains $b - 0$ if $b \geq 0$ and $b - ba$ if $b < 0$. Hence S contains a least element r. Then $b = qa + r$, where $q, r \in \mathbb{Z}$ and $r \geq 0$. Since $r - a = b - (q + 1)a$ and r is the least element in S, we must also have $r < a$. □

The concept of number will now be further extended to include 'fractions' or 'rational numbers'. For measuring lengths the integers do not suffice, since the length of a given segment may not be an exact multiple of the chosen unit of length. Similarly for measuring weights, if we find that three identical coins balance five of the chosen unit weights, then we ascribe to each coin the weight 5/3. In the realm of integers the equation $ax = b$ frequently has no solution; in the extended realm of rational numbers it will always have a solution if $a \neq 0$.

Intuitively, a rational number is the ratio or 'quotient' a/b of two integers a, b, where $b \neq 0$, with addition and multiplication defined by

$$a/b + c/d = (ad + cb)/bd,$$
$$a/b \cdot c/d = ac/bd.$$

However, two other integers a', b' may have the same ratio as a, b, and anyway what does a/b mean? To make things precise, we proceed in much the same way as before.

Put $\mathbb{Z}^\times = \mathbb{Z}\backslash\{0\}$ and consider the set $\mathbb{Z} \times \mathbb{Z}^\times$ of all ordered pairs (a, b) with $a \in \mathbb{Z}$ and $b \in \mathbb{Z}^\times$. For any two such ordered pairs, (a, b) and (a', b'), we write

$$(a, b) \sim (a', b') \quad \text{if } ab' = a'b.$$

To show that this is an equivalence relation it is again enough to verify that $(a, b) \sim (a', b')$ and $(a', b') \sim (a'', b'')$ imply $(a, b) \sim (a'', b'')$. The same calculation as before, with addition replaced by multiplication, shows that $(ab'')b' = (a''b)b'$. Since $b' \neq 0$, it follows that $ab'' = a''b$.

The equivalence class containing $(0, 1)$ evidently consists of all pairs $(0, b)$ with $b \neq 0$, and the equivalence class containing $(1, 1)$ consists of all pairs (b, b) with $b \neq 0$.

We define a *rational number* to be an equivalence class of elements of $\mathbb{Z} \times \mathbb{Z}^\times$ and, as is now customary, we denote the set of all rational numbers by $\mathbb{Q}$.

Addition of rational numbers is defined by

$$(a, b) + (c, d) = (ad + cb, bd),$$

where $bd \neq 0$ since $b \neq 0$ and $d \neq 0$. To justify the definition we must show that

$$(a, b) \sim (a', b') \text{ and } (c, d) \sim (c', d') \text{ imply } (ad + cb, bd) \sim (a'd' + c'b', b'd').$$

But if $ab' = a'b$ and $cd' = c'd$, then

$$\begin{aligned}(ad + cb)(b'd') &= (ab')(dd') + (cd')(bb') \\ &= (a'b)(dd') + (c'd)(bb') = (a'd' + c'b')(bd).\end{aligned}$$

It is easily verified that, also in $\mathbb{Q}$, addition satisfies the commutative law **(A2)** and the associative law **(A3)**. Moreover **(A4)** and **(A5)** also hold, the equivalence class 0 containing $(0, 1)$ being an identity element for addition and the equivalence class containing $(-b, c)$ being the additive inverse of the equivalence class containing (b, c).

Multiplication of rational numbers is defined componentwise:

$$(a, b) \cdot (c, d) = (ac, bd).$$

To justify the definition we must show that

$$(a, b) \sim (a', b') \text{ and } (c, d) \sim (c', d') \text{ imply } (ac, bd) \sim (a'c', b'd').$$

But if $ab' = a'b$ and $cd' = c'd$, then

$$(ac)(b'd') = (ab')(cd') = (a'b)(c'd) = (a'c')(bd).$$

It is easily verified that, also in $\mathbb{Q}$, multiplication satisfies the commutative law **(M2)** and the associative law **(M3)**. Moreover **(M4)** also holds, the equivalence class 1 containing $(1, 1)$ being an identity element for multiplication. Furthermore, addition and multiplication are connected by the distributive law **(AM1)**, and **(AM2)** also holds since $(0, 1)$ is not equivalent to $(1, 1)$.

Unlike the situation for $\mathbb{Z}$, however, every nonzero element of $\mathbb{Q}$ has a *multiplicative inverse*:

(M5) *for each* $a \neq 0$, *there exists* a^{-1} *such that* $aa^{-1} = 1$.

In fact, if a is represented by (b, c), then a^{-1} is represented by (c, b).

It follows that, for all $a, b \in \mathbb{Q}$ with $a \neq 0$, the equation $ax = b$ has a unique solution $x \in \mathbb{Q}$, namely $x = a^{-1}b$. Hence, if $a \neq 0$, then 1 is the only solution of $ax = a$, a^{-1} is uniquely determined by a, and $a = (a^{-1})^{-1}$.

We will say that a rational number a is *positive* if it is represented by an ordered pair (b, c) of integers for which $bc > 0$. This definition does not depend on the choice of representative. For suppose $0 < bc$ and $bc' = b'c$. Then $bc' \neq 0$, since $b \neq 0$ and $c' \neq 0$, and hence $0 < (bc')^2$. Since $(bc')^2 = (bc)(b'c')$ and $0 < bc$, it follows that $0 < b'c'$.

Our previous use of P having been abandoned in favour of $\mathbb{N}$, we will now denote by P the set of all positive rational numbers and by $-P$ the set of all rational numbers

$-a$, where $a \in P$. From the corresponding result for $\mathbb{Z}$, it follows that **(P1)** continues to hold in $\mathbb{Q}$. We will show that **(P2)** and **(P3)** also hold.

To see that the sum of two positive rational numbers is again positive, we observe that if a, b, c, d are integers such that $0 < ab$ and $0 < cd$, then also

$$0 < (ab)d^2 + (cd)b^2 = (ad + cb)(bd).$$

To see that the product of two positive rational numbers is again positive, we observe that if a, b, c, d are integers such that $0 < ab$ and $0 < cd$, then also

$$0 < (ab)(cd) = (ac)(bd).$$

Since **(P1)–(P3)** all hold, it follows as before that Propositions 12 and 13 also hold in $\mathbb{Q}$. Hence $1 \in P$ and **(O4)′** now implies that $a^{-1} \in P$ if $a \in P$. If $a, b \in P$ and $a < b$, then $b^{-1} < a^{-1}$, since $bb^{-1} = 1 = aa^{-1} < ba^{-1}$.

The set P of positive elements now induces an order relation on $\mathbb{Q}$. We write $a < b$ if $b - a \in P$, so that $a \in P$ if and only if $0 < a$. Then the order relations **(O1)–(O3)** and **(O4)′** continue to hold in $\mathbb{Q}$.

Unlike the situation for $\mathbb{Z}$, however, the ordering of $\mathbb{Q}$ is *dense*, i.e. if $a, b \in \mathbb{Q}$ and $a < b$, then there exists $c \in \mathbb{Q}$ such that $a < c < b$. For example, we can take c to be the solution of $(1 + 1)c = a + b$.

Let $\mathbb{Z}'$ denote the set of all rational numbers a' which can be represented by $(a, 1)$ for some $a \in \mathbb{Z}$. For every $c \in \mathbb{Q}$, there exist $a', b' \in \mathbb{Z}'$ with $b' \neq 0$ such that $c = a'b'^{-1}$. In fact, if c is represented by (a, b), we can take a' to be represented by $(a, 1)$ and b' by $(b, 1)$. Instead of $c = a'b'^{-1}$, we also write $c = a'/b'$.

For any $a \in \mathbb{Z}$, let a' be the rational number represented by $(a, 1)$. The map $a \to a'$ of $\mathbb{Z}$ into $\mathbb{Z}'$ is clearly bijective. Moreover, it preserves sums and products:

$$(a + b)' = a' + b', \quad (ab)' = a'b'.$$

Furthermore,

$$a' < b' \quad \text{if and only if } a < b.$$

Thus the map $a \to a'$ establishes an 'isomorphism' of $\mathbb{Z}$ with $\mathbb{Z}'$, and $\mathbb{Z}'$ is a copy of $\mathbb{Z}$ situated within $\mathbb{Q}$. By identifying a with a', we may regard $\mathbb{Z}$ itself as a subset of $\mathbb{Q}$. Then any rational number is the ratio of two integers.

By way of illustration, we show that if a and b are positive rational numbers, then there exists a positive integer l such that $la > b$. For if $a = m/n$ and $b = p/q$, where m, n, p, q are positive integers, then

$$(np + 1)a > pm \geq p \geq b.$$

3 Real Numbers

It was discovered by the ancient Greeks that even rational numbers do not suffice for the measurement of lengths. If x is the length of the hypotenuse of a right-angled triangle whose other two sides have unit length then, by Pythagoras' theorem, $x^2 = 2$.

But it was proved, probably by a disciple of Pythagoras, that there is no rational number x such that $x^2 = 2$. (A more general result is proved in Book X, Proposition 9 of Euclid's *Elements*.) We give here a somewhat different proof from the classical one.

Assume that such a rational number x exists. Since x may be replaced by $-x$, we may suppose that $x = m/n$, where $m, n \in \mathbb{N}$. Then $m^2 = 2n^2$. Among all pairs m, n of positive integers with this property, there exists one for which n is least. If we put

$$p = 2n - m, \quad q = m - n,$$

then p and q are positive integers, since clearly $n < m < 2n$. But

$$p^2 = 4n^2 - 4mn + m^2 = 2(m^2 - 2mn + n^2) = 2q^2.$$

Since $q < n$, this contradicts the minimality of n.

If we think of the rational numbers as measuring distances of points on a line from a given origin O on the line (with distances on one side of O positive and distances on the other side negative), this means that, even though a dense set of points is obtained in this way, not all points of the line are accounted for. In order to fill in the gaps the concept of number will now be extended from 'rational number' to 'real number'.

It is possible to define real numbers as infinite decimal expansions, the rational numbers being those whose decimal expansions are eventually periodic. However, the choice of base 10 is arbitrary and carrying through this approach is awkward.

There are two other commonly used approaches, one based on *order* and the other on *distance*. The first was proposed by Dedekind (1872), the second by Méray (1869) and Cantor (1872). We will follow Dedekind's approach, since it is conceptually simpler. However, the second method is also important and in a sense more general. In Chapter VI we will use it to extend the rational numbers to the *p-adic numbers*.

It is convenient to carry out Dedekind's construction in two stages. We will first define 'cuts' (which are just the positive real numbers), and then pass from cuts to arbitrary real numbers in the same way that we passed from the natural numbers to the integers.

Intuitively, a cut is the set of all rational numbers which represent points of the line between the origin O and some other point. More formally, we define a *cut* to be a nonempty proper subset A of the set P of all positive rational numbers such that

(i) *if* $a \in A, b \in P$ *and* $b < a$, *then* $b \in A$;
(ii) *if* $a \in A$, *then there exists* $a' \in A$ *such that* $a < a'$.

For example, the set I of all positive rational numbers $a < 1$ is a cut. Similarly, the set T of all positive rational numbers a such that $a^2 < 2$ is a cut. We will denote the set of all cuts by $\mathscr{P}$.

For any $A, B \in \mathscr{P}$ we write $A < B$ if A is a proper subset of B. We will show that this induces a *total order* on $\mathscr{P}$.

It is clear that if $A < B$ and $B < C$, then $A < C$. It remains to show that, for any $A, B \in \mathscr{P}$, one and only one of the following alternatives holds:

$$A < B, \quad A = B, \quad B < A.$$

It is obvious from the definition by set inclusion that at most one holds. Now suppose that neither $A < B$ nor $A = B$. Then there exists $a \in A \backslash B$. It follows from (i), applied to B, that every $b \in B$ satisfies $b < a$ and then from (i), applied to A, that $b \in A$. Thus $B < A$.

Let $\mathscr{S}$ be any nonempty collection of cuts. A cut B is said to be an *upper bound* for $\mathscr{S}$ if $A \leq B$ for every $A \in \mathscr{S}$, and a *lower bound* for $\mathscr{S}$ if $B \leq A$ for every $A \in \mathscr{S}$. An upper bound for $\mathscr{S}$ is said to be a *least upper bound* or *supremum* for $\mathscr{S}$ if it is a lower bound for the collection of all upper bounds. Similarly, a lower bound for $\mathscr{S}$ is said to be a *greatest lower bound* or *infimum* for $\mathscr{S}$ if it is an upper bound for the collection of all lower bounds. Clearly, $\mathscr{S}$ has at most one supremum and at most one infimum.

The set $\mathscr{P}$ has the following basic property:

(P4) *if a nonempty subset $\mathscr{S}$ has an upper bound, then it has a least upper bound.*

Proof Let C be the union of all sets $A \in \mathscr{S}$. By hypothesis there exists a cut B such that $A \subseteq B$ for every $A \in \mathscr{S}$. Since $C \subseteq B$ for any such B, and $A \subseteq C$ for every $A \in \mathscr{S}$, we need only show that C is a cut.

Evidently C is a nonempty proper subset of P, since $B \neq P$. Suppose $c \in C$. Then $c \in A$ for some $A \in \mathscr{S}$. If $d \in P$ and $d < c$, then $d \in A$, since A is a cut. Furthermore $c < a'$ for some $a' \in A$. Since $A \subseteq C$, this proves that C is a cut. □

In the set P of positive rational numbers, the subset T of all $x \in P$ such that $x^2 < 2$ has an upper bound, but no least upper bound. Thus **(P4)** shows that there is a difference between the total order on P and that on $\mathscr{P}$.

We now define addition of cuts. For any $A, B \in \mathscr{P}$, let $A + B$ denote the set of all rational numbers $a + b$, with $a \in A$ and $b \in B$. We will show that also $A + B \in \mathscr{P}$. Evidently $A + B$ is a nonempty subset of P. It is also a proper subset. For choose $c \in P \backslash A$ and $d \in P \backslash B$. Then, by (i), $a < c$ for all $a \in A$ and $b < d$ for all $b \in B$. Since $a + b < c + d$ for all $a \in A$ and $b \in B$, it follows that $c + d \notin A + B$.

Suppose now that $a \in A$, $b \in B$ and that $c \in P$ satisfies $c < a + b$. If $c > b$, then $c = b + d$ for some $d \in P$, and $d < a$. Hence, by (i), $d \in A$ and $c = d + b \in A + B$. Similarly, $c \in A + B$ if $c > a$. Finally, if $c \leq a$ and $c \leq b$, choose $e \in P$ so that $e < c$. Then $e \in A$ and $c = e + f$ for some $f \in P$. Then $f \in B$, since $f < c$, and $c = e + f \in A + B$.

Thus $A + B$ has the property (i). It is trivial that $A + B$ also has the property (ii), since if $a \in A$ and $b \in B$, there exists $a' \in A$ such that $a < a'$ and then $a + b < a' + b$. This completes the proof that $A + B$ is a cut.

It follows at once from the corresponding properties of rational numbers that addition of cuts satisfies the commutative law **(A2)** and the associative law **(A3)**.

We consider next the connection between addition and order.

Lemma 15 *For any cut A and any $c \in P$, there exists $a \in A$ such that $a + c \notin A$.*

Proof If $c \notin A$, then $a + c \notin A$ for every $a \in A$, since $c < a + c$. Thus we may suppose $c \in A$. Choose $b \in P \backslash A$. For some positive integer n we have $b < nc$ and hence $nc \notin A$. If n is the least positive integer such that $nc \notin A$, then $n > 1$ and $(n - 1)c \in A$. Consequently we can take $a = (n - 1)c$. □

Proposition 16 *For any cuts A, B, there exists a cut C such that $A + C = B$ if and only if $A < B$.*

Proof We prove the necessity of the condition by showing that $A < A + C$ for any cuts A, C. If $a \in A$ and $c \in C$, then $a < a + c$. Since $A + C$ is a cut, it follows that $a \in A + C$. Consequently $A \leq A + C$, and Lemma 15 implies that $A \neq A + C$.

Suppose now that A and B are cuts such that $A < B$, and let C be the set of all $c \in P$ such that $c + d \in B$ for some $d \in P \backslash A$. We are going to show that C is a cut and that $A + C = B$.

The set C is not empty. For choose $b \in B \backslash A$ and then $b' \in B$ with $b < b'$. Then $b' = b + c'$ for some $c' \in P$, which implies $c' \in C$. On the other hand, $C \leq B$, since $c + d \in B$ and $d \in P$ imply $c \in B$. Thus C is a proper subset of P.

Suppose $c \in C$, $p \in P$ and $p < c$. We have $c + d \in B$ for some $d \in P \backslash A$ and $c = p + e$ for some $e \in P$. Since $d + e \in P \backslash A$ and $p + (d + e) = c + d \in B$, it follows that $p \in C$.

Suppose now that $c \in C$, so that $c + d \in B$ for some $d \in P \backslash A$. Choose $b \in B$ so that $c + d < b$. Then $b = c + d + e$ for some $e \in P$. If we put $c' = c + e$, then $c < c'$. Moreover $c' \in C$, since $c' + d = b$. This completes the proof that C is a cut.

Suppose $a \in A$ and $c \in C$. Then $c + d \in B$ for some $d \in P \backslash A$. Hence $a < d$. It follows that $a + c < c + d$, and so $a + c \in B$. Thus $A + C \leq B$.

It remains to show that $B \leq A + C$. Pick any $b \in B$. If $b \in A$, then also $b \in A + C$, since $A < A + C$. Thus we now assume $b \notin A$. Choose $b' \in B$ with $b < b'$. Then $b' = b + d$ for some $d \in P$. By Lemma 15, there exists $a \in A$ such that $a + d \notin A$. Moreover $a < b$, since $b \notin A$, and hence $b = a + c$ for some $c \in P$. Since $c + (a + d) = b + d = b'$, it follows that $c \in C$. Thus $b \in A + C$ and $B \leq A + C$. □

We can now show that addition of cuts satisfies the order relation **(O3)**. Suppose first that $A < B$. Then, by Proposition 16, there exists a cut D such that $A + D = B$. Hence, for any cut C,

$$A + C < (A + C) + D = B + C.$$

Suppose next that $A + C < B + C$. Then $A \neq B$. Since $B < A$ would imply $B + C < A + C$, by what we have just proved, it follows from the law of trichotomy that $A < B$.

From **(O3)** and the law of trichotomy, it follows that addition of cuts satisfies the cancellation law **(A1)**.

We next define multiplication of cuts. For any $A, B \in \mathscr{P}$, let AB denote the set of all rational numbers ab, with $a \in A$ and $b \in B$. In the same way as for $A + B$, it may be shown that $AB \in \mathscr{P}$. We note only that if $a \in A, b \in B$ and $c < ab$, then $b^{-1}c < a$. Hence $b^{-1}c \in A$ and $c = (b^{-1}c)b \in AB$.

It follows from the corresponding properties of rational numbers that multiplication of cuts satisfies the commutative law **(M2)** and the associative law **(M3)**. Moreover **(M4)** holds, the identity element for multiplication being the cut I consisting of all positive rational numbers less than 1.

We now show that the distributive law **(AM1)** also holds. The distributive law for rational numbers shows at once that

$$A(B + C) \leq AB + AC.$$

It remains to show that $a_1b + a_2c \in A(B + C)$ if $a_1, a_2 \in A$, $b \in B$ and $c \in C$. But

$$a_1b + a_2c \leq a_2(b + c) \quad \text{if } a_1 \leq a_2,$$

and

$$a_1b + a_2c \leq a_1(b + c) \quad \text{if } a_2 \leq a_1.$$

In either event it follows that $a_1b + a_2c \in A(B + C)$.

We can now show that multiplication of cuts satisfies the order relation **(O4)**. If $A < B$, then there exists a cut D such that $A + D = B$ and hence $AC < AC + DC = BC$. Conversely, suppose $AC < BC$. Then $A \neq B$. Since $B < A$ would imply $BC < AC$, it follows that $A < B$.

From **(O4)** and the law of trichotomy **(O2)** it follows that multiplication of cuts satisfies the cancellation law **(M1)**.

We next prove the existence of multiplicative inverses. The proof will use the following multiplicative analogue of Lemma 15:

Lemma 17 *For any cut A and any $c \in P$ with $c > 1$, there exists $a \in A$ such that $ac \notin A$.*

Proof Choose any $b \in A$. We may suppose $bc \in A$, since otherwise we can take $a = b$. Since $b < bc$, we have $bc = b + d$ for some $d \in P$. By Lemma 15 we can choose $a \in A$ so that $a + d \notin A$. Since $b + d \in A$, it follows that $b + d < a + d$, and so $b < a$. Hence $ab^{-1} > 1$ and

$$a + d < a + (ab^{-1})d = ab^{-1}(b + d) = ac.$$

Since $a + d \notin A$, it follows that $ac \notin A$. □

Proposition 18 *For any $A \in \mathscr{P}$, there exists $A^{-1} \in \mathscr{P}$ such that $AA^{-1} = I$.*

Proof Let A^{-1} be the set of all $b \in P$ such that $b < c^{-1}$ for some $c \in P \backslash A$. It is easily verified that A^{-1} is a cut. We note only that $a^{-1} \notin A^{-1}$ if $a \in A$ and that, if $b < c^{-1}$, then also $b < d^{-1}$ for some $d > c$.

We now show that $AA^{-1} = I$. If $a \in A$ and $b \in A^{-1}$ then $ab < 1$, since $a \geq b^{-1}$ would imply $a > c$ for some $c \in P \backslash A$. Thus $AA^{-1} \leq I$. On the other hand, if $0 < d < 1$ then, by Lemma 17, there exists $a \in A$ such that $ad^{-1} \notin A$. Choose $a' \in A$ so that $a < a'$, and put $b = (a')^{-1}d$. Then $b < a^{-1}d$. Since $a^{-1}d = (ad^{-1})^{-1}$, it follows that $b \in A^{-1}$ and consequently $d = a'b \in AA^{-1}$. Thus $I \leq AA^{-1}$. □

For any positive rational number a, the set A_a consisting of all positive rational numbers c such that $c < a$ is a cut. The map $a \to A_a$ of P into $\mathscr{P}$ is injective and preserves sums and products:

$$A_{a+b} = A_a + A_b, A_{ab} = A_aA_b.$$

Moreover, $A_a < A_b$ if and only if $a < b$.

By identifying a with A_a we may regard P as a subset of $\mathscr{P}$. It is a proper subset, since **(P4)** does not hold in P.

This completes the first stage of Dedekind's construction. In the second stage we pass from cuts to real numbers. Intuitively, a real number is the difference of two cuts. We will deal with the second stage rather briefly since, as has been said, it is completely analogous to the passage from the natural numbers to the integers.

On the set $\mathscr{P} \times \mathscr{P}$ of all ordered pairs of cuts an equivalence relation is defined by

$$(A, B) \sim (A', B') \quad \text{if } A + B' = A' + B.$$

We define a *real number* to be an equivalence class of ordered pairs of cuts and, as is now customary, we denote the set of all real numbers by $\mathbb{R}$.

Addition and multiplication are unambiguously defined by

$$(A, B) + (C, D) = (A + C, B + D),$$
$$(A, B) \cdot (C, D) = (AC + BD, AD + BC).$$

They obey the laws **(A2)–(A5)**, **(M2)–(M5)** and **(AM1)–(AM2)**.

A real number represented by (A, B) is said to be *positive* if $B < A$. If we denote by $\mathscr{P}'$ the set of all positive real numbers, then **(P1)–(P3)** continue to hold with $\mathscr{P}'$ in place of P. An order relation, satisfying **(O1)–(O3)**, is induced on $\mathbb{R}$ by writing $a < b$ if $b - a \in \mathscr{P}'$. Moreover, any $a \in \mathbb{R}$ may be written in the form $a = b - c$, where $b, c \in \mathscr{P}'$. It is easily seen that $\mathscr{P}$ is isomorphic with $\mathscr{P}'$. By identifying $\mathscr{P}$ with $\mathscr{P}'$, we may regard both $\mathscr{P}$ and $\mathbb{Q}$ as subsets of $\mathbb{R}$. An element of $\mathbb{R}\backslash\mathbb{Q}$ is said to be an *irrational* real number.

Upper and lower bounds, and suprema and infima, may be defined for subsets of $\mathbb{R}$ in the same way as for subsets of $\mathscr{P}$. Moreover, the least upper bound property **(P4)** continues to hold in $\mathbb{R}$. By applying **(P4)** to the subset $-\mathscr{S} = \{-a : a \in \mathscr{S}\}$ we see that if a nonempty subset $\mathscr{S}$ of $\mathbb{R}$ has a lower bound, then it has a greatest lower bound.

The least upper bound property implies the so-called *Archimedean property*:

Proposition 19 *For any positive real numbers a, b, there exists a positive integer n such that $na > b$.*

Proof Assume, on the contrary, that $na \leq b$ for every $n \in \mathbb{N}$. Then b is an upper bound for the set $\{na : n \in \mathbb{N}\}$. Let c be a least upper bound for this set. From $na \leq c$ for every $n \in \mathbb{N}$ we obtain $(n + 1)a \leq c$ for every $n \in \mathbb{N}$. But this implies $na \leq c - a$ for every $n \in \mathbb{N}$. Since $c - a < c$ and c is a least upper bound, we have a contradiction. □

Proposition 20 *For any real numbers a, b with $a < b$, there exists a rational number c such that $a < c < b$.*

Proof Suppose first that $a \geq 0$. By Proposition 19 there exists a positive integer n such that $n(b - a) > 1$. Then $b > a + n^{-1}$. There exists also a positive integer m such that $mn^{-1} > a$. If m is the least such positive integer, then $(m - 1)n^{-1} \leq a$ and hence $mn^{-1} \leq a + n^{-1} < b$. Thus we can take $c = mn^{-1}$.

If $a < 0$ and $b > 0$ we can take $c = 0$. If $a < 0$ and $b \leq 0$, then $-b < d < -a$ for some rational d and we can take $c = -d$. □

Proposition 21 *For any positive real number a, there exists a unique positive real number b such that $b^2 = a$.*

Proof Let S be the set of all positive real numbers x such that $x^2 \leq a$. The set S is not empty, since it contains a if $a \leq 1$ and 1 if $a > 1$. If $y > 0$ and $y^2 > a$, then y is an upper bound for S. In particular, $1 + a$ is an upper bound for S. Let b be the least upper bound for S. Then $b^2 = a$, since $b^2 < a$ would imply $(b + 1/n)^2 < a$ for sufficiently large $n > 0$ and $b^2 > a$ would imply $(b - 1/n)^2 > a$ for sufficiently large $n > 0$. Finally, if $c^2 = a$ and $c > 0$, then $c = b$, since

$$(c - b)(c + b) = c^2 - b^2 = 0. \qquad \square$$

The unique positive real number b in the statement of Proposition 21 is said to be a *square root* of a and is denoted by $\sqrt{a}$ or $a^{1/2}$. In the same way it may be shown that, for any positive real number a and any positive integer n, there exists a unique positive real number b such that $b^n = a$, where $b^n = b \cdots b$ (n times). We say that b is an *n-th root* of a and write $b = \sqrt[n]{a}$ or $a^{1/n}$.

A set is said to be a *field* if two binary operations, addition and multiplication, are defined on it with the properties **(A2)–(A5)**, **(M2)–(M5)** and **(AM1)–(AM2)**. A field is said to be *ordered* if it contains a subset P of 'positive' elements with the properties **(P1)–(P3)**. An ordered field is said to be *complete* if, with the order induced by P, it has the property **(P4)**.

Propositions 19–21 hold in any complete ordered field, since only the above properties were used in their proofs. By construction, the set $\mathbb{R}$ of all real numbers is a complete ordered field. In fact, any complete ordered field F is isomorphic to $\mathbb{R}$, i.e. there exists a bijective map $\varphi : F \to \mathbb{R}$ such that, for all $a, b \in F$,

$$\varphi(a + b) = \varphi(a) + \varphi(b),$$
$$\varphi(ab) = \varphi(a)\varphi(b),$$

and $\varphi(a) > 0$ if and only if $a \in P$. We sketch the proof.

Let e be the identity element for multiplication in F and, for any positive integer n, let $ne = e + \cdots + e$ (n summands). Since F is ordered, ne is positive and so has a multiplicative inverse. For any rational number m/n, where $m, n \in \mathbb{Z}$ and $n > 0$, write $(m/n)e = m(ne)^{-1}$ if $m > 0$, $= -(-m)(ne)^{-1}$ if $m < 0$, and $= 0$ if $m = 0$. The elements $(m/n)e$ form a subfield of F isomorphic to $\mathbb{Q}$ and we define $\varphi((m/n)e) = m/n$. For any $a \in F$, we define $\varphi(a)$ to be the least upper bound of all rational numbers m/n such that $(m/n)e \leq a$. One verifies first that the map $\varphi : F \to \mathbb{R}$ is bijective and that $\varphi(a) < \varphi(b)$ if and only if $a < b$. One then deduces that φ preserves sums and products.

Actually, any bijective map $\varphi : F \to \mathbb{R}$ which preserves sums and products is also order-preserving. For, by Proposition 21, $b > a$ if and only if $b - a = c^2$ for some $c \neq 0$, and then

$$\varphi(b) - \varphi(a) = \varphi(b - a) = \varphi(c^2) = \varphi(c)^2 > 0.$$

Those whose primary interest lies in real analysis may *define* $\mathbb{R}$ to be a complete ordered field and omit the tour through $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ and $\mathscr{P}$. That is, one takes as axioms the 14 properties above which define a complete ordered field and simply assumes that they are consistent.

The notion of convergence can be defined in any totally ordered set. A sequence $\{a_n\}$ is said to *converge*, with *limit* l, if for any l', l'' such that $l' < l < l''$, there exists a positive integer $N = N(l', l'')$ such that

$$l' < a_n < l'' \quad \text{for every } n \geq N.$$

The limit l of the convergent sequence $\{a_n\}$ is clearly uniquely determined; we write

$$\lim_{n\to\infty} a_n = l,$$

or $a_n \to l$ as $n \to \infty$.

It is easily seen that any convergent sequence is *bounded*, i.e. it has an upper bound and a lower bound. A trivial example of a convergent sequence is the *constant* sequence $\{a_n\}$, where $a_n = a$ for every n; its limit is again a.

In the set $\mathbb{R}$ of real numbers, or in any totally ordered set in which each bounded sequence has a least upper bound and a greatest lower bound, the definition of convergence can be reformulated. For, let $\{a_n\}$ be a bounded sequence. Then, for any positive integer m, the subsequence $\{a_n\}_{n\geq m}$ has a greatest lower bound b_m and a least upper bound c_m:

$$b_m = \inf_{n\geq m} a_n, \quad c_m = \sup_{n\geq m} a_n.$$

The sequences $\{b_m\}_{m\geq 1}$ and $\{c_m\}_{m\geq 1}$ are also bounded and, for any positive integer m,

$$b_m \leq b_{m+1} \leq c_{m+1} \leq c_m.$$

If we define the *lower limit* and *upper limit* of the sequence $\{a_n\}$ by

$$\underline{\lim_{n\to\infty}}\, a_n := \sup_{m\geq 1} b_m, \quad \overline{\lim_{n\to\infty}}\, a_n := \inf_{m\geq 1} c_m,$$

then $\underline{\lim}_{n\to\infty} a_n \leq \overline{\lim}_{n\to\infty} a_n$, and it is readily shown that $\lim_{n\to\infty} a_n = l$ if and only if

$$\underline{\lim_{n\to\infty}}\, a_n = l = \overline{\lim_{n\to\infty}}\, a_n.$$

A sequence $\{a_n\}$ is said to be *nondecreasing* if $a_n \leq a_{n+1}$ for every n and *nonincreasing* if $a_{n+1} \leq a_n$ for every n. It is said to be *monotonic* if it is either nondecreasing or nonincreasing.

Proposition 22 *Any bounded monotonic sequence of real numbers is convergent.*

Proof Let $\{a_n\}$ be a bounded monotonic sequence and suppose, for definiteness, that it is nondecreasing: $a_1 \leq a_2 \leq a_3 \leq \cdots$. In this case, in the notation used above we have $b_m = a_m$ and $c_m = c_1$ for every m. Hence

$$\lim_{n\to\infty} a_n = \sup_{m\geq 1} a_m = c_1 = \overline{\lim_{n\to\infty}}\, a_n. \qquad \square$$

Proposition 22 may be applied to the centuries-old algorithm for calculating square roots, which is commonly used today in pocket calculators. Take any real number $a > 1$ and put

$$x_1 = (1 + a)/2.$$

Then $x_1 > 1$ and $x_1^2 > a$, since $(a-1)^2 > 0$. Define the sequence $\{x_n\}$ recursively by

$$x_{n+1} = (x_n + a/x_n)/2 \quad (n \geq 1).$$

It is easily verified that if $x_n > 1$ and $x_n^2 > a$, then $x_{n+1} > 1$, $x_{n+1}^2 > a$ and $x_{n+1} < x_n$. Since the inequalities hold for $n = 1$, it follows that they hold for all n. Thus the sequence $\{x_n\}$ is nonincreasing and bounded, and therefore convergent. If $x_n \to b$, then $a/x_n \to a/b$ and $x_{n+1} \to b$. Hence $b = (b + a/b)/2$, which simplifies to $b^2 = a$.

We consider now sequences of real numbers which are not necessarily monotonic.

Lemma 23 *Any sequence $\{a_n\}$ of real numbers has a monotonic subsequence.*

Proof Let M be the set of all positive integers m such that $a_m \geq a_n$ for every $n > m$. If M contains infinitely many positive integers $m_1 < m_2 < \cdots$, then $\{a_{m_k}\}$ is a nonincreasing subsequence of $\{a_n\}$. If M is empty or finite, there is a positive integer n_1 such that no positive integer $n \geq n_1$ is in M. Then $a_{n_2} > a_{n_1}$ for some $n_2 > n_1$, $a_{n_3} > a_{n_2}$ for some $n_3 > n_2$, and so on. Thus $\{a_{n_k}\}$ is a nondecreasing subsequence of $\{a_n\}$. □

It is clear from the proof that Lemma 23 also holds for sequences of elements of any totally ordered set. In the case of $\mathbb{R}$, however, it follows at once from Lemma 23 and Proposition 22 that

Proposition 24 *Any bounded sequence of real numbers has a convergent subsequence.*

Proposition 24 is often called the Bolzano–Weierstrass theorem. It was stated by Bolzano (c. 1830) in work which remained unpublished until a century later. It became generally known through the lectures of Weierstrass (c. 1874).

A sequence $\{a_n\}$ of real numbers is said to be a *fundamental sequence*, or 'Cauchy sequence', if for each $\varepsilon > 0$ there exists a positive integer $N = N(\varepsilon)$ such that

$$-\varepsilon < a_p - a_q < \varepsilon \quad \text{for all } p, q \geq N.$$

Any fundamental sequence $\{a_n\}$ is bounded, since any finite set is bounded and

$$a_N - \varepsilon < a_p < a_N + \varepsilon \quad \text{for } p \geq N.$$

Also, any convergent sequence is a fundamental sequence. For suppose $a_n \to l$ as $n \to \infty$. Then, for any $\varepsilon > 0$, there exists a positive integer N such that

$$l - \varepsilon/2 < a_n < l + \varepsilon/2 \quad \text{for every } n \geq N.$$

It follows that

$$-\varepsilon < a_p - a_q < \varepsilon \quad \text{for } p \geq q \geq N.$$

The definitions of convergent sequence and fundamental sequence, and the preceding result that 'convergent' implies 'fundamental', hold also for sequences of rational numbers, and even for sequences with elements from any ordered field. However, for sequences of real numbers there is a converse result:

Proposition 25 *Any fundamental sequence of real numbers is convergent.*

Proof If $\{a_n\}$ is a fundamental sequence of real numbers, then $\{a_n\}$ is bounded and, for any $\varepsilon > 0$, there exists a positive integer $m = m(\varepsilon)$ such that

$$-\varepsilon/2 < a_p - a_q < \varepsilon/2 \quad \text{for all } p, q \geq m.$$

But, by Proposition 24, the sequence $\{a_n\}$ has a convergent subsequence $\{a_{n_k}\}$. If l is the limit of this subsequence, then there exists a positive integer $N \geq m$ such that

$$l - \varepsilon/2 < a_{n_k} < l + \varepsilon/2 \quad \text{for } n_k \geq N.$$

It follows that

$$l - \varepsilon < a_n < l + \varepsilon \quad \text{for } n \geq N.$$

Thus the sequence $\{a_n\}$ converges with limit l. □

Proposition 25 was known to Bolzano (1817) and was clearly stated in the influential *Cours d'analyse* of Cauchy (1821). However, a rigorous proof was impossible until the real numbers themselves had been precisely defined.

The Méray–Cantor method of constructing the real numbers from the rationals is based on Proposition 25. We define two fundamental sequences $\{a_n\}$ and $\{a'_n\}$ of rational numbers to be equivalent if $a_n - a'_n \to 0$ as $n \to \infty$. This is indeed an equivalence relation, and we define a real number to be an equivalence class of fundamental sequences. The set of all real numbers acquires the structure of a field if addition and multiplication are defined by

$$\{a_n\} + \{b_n\} = \{a_n + b_n\}, \quad \{a_n\} \cdot \{b_n\} = \{a_n b_n\}.$$

It acquires the structure of a complete ordered field if the fundamental sequence $\{a_n\}$ is said to be positive when it has a positive lower bound. The field $\mathbb{Q}$ of rational numbers may be regarded as a subfield of the field thus constructed by identifying the rational number a with the equivalence class containing the constant sequence $\{a_n\}$, where $a_n = a$ for every n.

It is not difficult to show that an ordered field is complete if every bounded monotonic sequence is convergent, or if every bounded sequence has a convergent subsequence. In this sense, Propositions 22 and 24 state equivalent forms for the least upper bound property. This is not true, however, for Proposition 25. An ordered field need not have the least upper bound property, even though every fundamental sequence is convergent. It is true, however, that an ordered field has the least upper bound property if and only if it has the Archimedean property (Proposition 19) *and* every fundamental sequence is convergent.

In a course of real analysis one would now define continuity and prove those properties of continuous functions which, in the 18th century, were assumed as 'geometrically obvious'. For example, for given $a, b \in \mathbb{R}$ with $a < b$, let $I = [a, b]$ be the *interval* consisting of all $x \in \mathbb{R}$ such that $a \leq x \leq b$. If $f : I \to \mathbb{R}$ is continuous, then it attains its supremum, i.e. there exists $c \in I$ such that $f(x) \leq f(c)$ for every $x \in I$. Also, if $f(a)f(b) < 0$, then $f(d) = 0$ for some $d \in I$ (the intermediate-value theorem). Real analysis is not our primary concern, however, and we do not feel obliged to establish even those properties which we may later use.

4 Metric Spaces

The notion of convergence is meaningful not only for points on a line, but also for points in space, where there is no natural relation of order. We now reformulate our previous definition, so as to make it more generally applicable.

The *absolute value* $|a|$ of a real number a is defined by

$$|a| = a \quad \text{if } a \geq 0,$$
$$|a| = -a \quad \text{if } a < 0.$$

It is easily seen that absolute values have the following properties:

$$|0| = 0, |a| > 0 \quad \text{if } a \neq 0;$$
$$|a| = |-a|;$$
$$|a + b| \leq |a| + |b|.$$

The first two properties follow at once from the definition. To prove the third, we observe first that $a + b \leq |a| + |b|$, since $a \leq |a|$ and $b \leq |b|$. Replacing a by $-a$ and b by $-b$, we obtain also $-(a + b) \leq |a| + |b|$. But $|a + b|$ is either $a + b$ or $-(a + b)$.

The *distance* between two real numbers a and b is defined to be the real number

$$\mathrm{d}(a, b) = |a - b|.$$

From the preceding properties of absolute values we obtain their counterparts for distances:

(D1) $\mathrm{d}(a, a) = 0, \mathrm{d}(a, b) > 0$ *if* $a \neq b$;
(D2) $\mathrm{d}(a, b) = \mathrm{d}(b, a)$;
(D3) $\mathrm{d}(a, b) \leq \mathrm{d}(a, c) + \mathrm{d}(c, b)$.

The third property is known as the *triangle inequality*, since it may be interpreted as saying that, in any triangle, the length of one side does not exceed the sum of the lengths of the other two.

Fréchet (1906) recognized these three properties as the essential characteristics of any measure of distance and introduced the following general concept. A set E is a *metric space* if with each ordered pair (a, b) of elements of E there is associated a real number $\mathrm{d}(a, b)$, so that the properties **(D1)–(D3)** hold for all $a, b, c \in E$.

We note first some simple consequences of these properties. For all $a, b, a', b' \in E$ we have

$$|\mathrm{d}(a, b) - \mathrm{d}(a', b')| \leq \mathrm{d}(a, a') + \mathrm{d}(b, b') \qquad (*)$$

since, by **(D2)** and **(D3)**,

$$\mathrm{d}(a, b) \leq \mathrm{d}(a, a') + \mathrm{d}(a', b') + \mathrm{d}(b, b'),$$
$$\mathrm{d}(a', b') \leq \mathrm{d}(a, a') + \mathrm{d}(a, b) + \mathrm{d}(b, b').$$

Taking $b = b'$ in $(*)$, we obtain from **(D1)**,

$$|\mathrm{d}(a, b) - \mathrm{d}(a', b)| \leq \mathrm{d}(a, a'). \qquad (**)$$

In any metric space there is a natural *topology*. A subset G of a metric space E is *open* if for each $x \in G$ there is a positive real number $\delta = \delta(x)$ such that G also contains the whole open ball $\beta_\delta(x) = \{y \in E : \mathrm{d}(x, y) < \delta\}$. A set $F \subseteq E$ is *closed* if its complement $E \setminus F$ is open.

For any set $A \subseteq E$, its *closure* $\bar{A}$ is the intersection of all closed sets containing it, and its *interior* $\mathrm{int}\, A$ is the union of all open sets contained in it.

A subset F of E is *connected* if it is not contained in the union of two open subsets of E whose intersections with F are disjoint and nonempty. A subset F of E is (sequentially) *compact* if every sequence of elements of F has a subsequence converging to an element of F (and *locally compact* if this holds for every bounded sequence of elements of F).

A map $f : X \to Y$ from one metric space X to another metric space Y is *continuous* if, for each open subset G of Y, the set of all $x \in X$ such that $f(x) \in G$ is an open subset of X. The two properties stated at the end of §3 admit far-reaching generalizations for continuous maps between subsets of metric spaces, namely that under a continuous map the image of a compact set is again compact, and the image of a connected set is again connected.

There are many examples of metric spaces:

(i) Let $E = \mathbb{R}^n$ be the set of all n-tuples $a = (\alpha_1, \ldots, \alpha_n)$ of real numbers and define

$$\mathrm{d}(b, c) = |b - c|,$$

where $b - c = (\beta_1 - \gamma_1, \ldots, \beta_n - \gamma_n)$ if $b = (\beta_1, \ldots, \beta_n)$ and $c = (\gamma_1, \ldots, \gamma_n)$, and

$$|a| = \max_{1 \le j \le n} |\alpha_j|.$$

Alternatively, one can replace the *norm* $|a|$ by either

$$|a|_1 = \sum_{j=1}^{n} |\alpha_j|$$

or

$$|a|_2 = \left(\sum_{j=1}^{n} |\alpha_j|^2 \right)^{1/2}.$$

In the latter case, $\mathrm{d}(b, c)$ is the *Euclidean distance* between b and c. The triangle inequality in this case follows from the *Cauchy–Schwarz inequality*: for any real numbers $\beta_j, \gamma_j (j = 1, \ldots, n)$

$$\left(\sum_{j=1}^{n} \beta_j \gamma_j \right)^2 \le \left(\sum_{j=1}^{n} \beta_j^2 \right) \left(\sum_{j=1}^{n} \gamma_j^2 \right).$$

(ii) Let $E = \mathbb{F}_2^n$ be the set of all n-tuples $a = (\alpha_1, \ldots, \alpha_n)$, where $\alpha_j = 0$ or 1 for each j, and define the *Hamming distance* $\mathrm{d}(b, c)$ between $b = (\beta_1, \ldots, \beta_n)$ and $c = (\gamma_1, \ldots, \gamma_n)$ to be the number of j such that $\beta_j \neq \gamma_j$. This metric space plays a basic role in the theory of *error-correcting codes*.

(iii) Let $E = \mathscr{C}(I)$ be the set of all continuous functions $f: I \to \mathbb{R}$, where

$$I = [a, b] = \{x \in \mathbb{R}: a \le x \le b\}$$

is an interval of $\mathbb{R}$, and define $\mathrm{d}(g, h) = |g - h|$, where

$$|f| = \sup_{a \le x \le b} |f(x)|.$$

(A well-known property of continuous functions ensures that f is bounded on I.) Alternatively, one can replace the *norm* $|f|$ by either

$$|f|_1 = \int_a^b |f(x)|dx$$

or

$$|f|_2 = \left(\int_a^b |f(x)|^2 dx\right)^{1/2}.$$

(iv) Let $E = \mathscr{C}(\mathbb{R})$ be the set of all continuous functions $f: \mathbb{R} \to \mathbb{R}$ and define

$$\mathrm{d}(g, h) = \sum_{N \ge 1} \mathrm{d}_N(g, h)/2^N[1 + \mathrm{d}_N(g, h)],$$

where $\mathrm{d}_N(g, h) = \sup_{|x| \le N} |g(x) - h(x)|$. The triangle inequality **(D3)** follows from the inequality

$$|\alpha + \beta|/[1 + |\alpha + \beta|] \le |\alpha|/[1 + |\alpha|] + |\beta|/[1 + |\beta|]$$

for arbitrary real numbers α, β.

The metric here has the property that $\mathrm{d}(f_n, f) \to 0$ if and only if $f_n(x) \to f(x)$ uniformly on every bounded subinterval of $\mathbb{R}$. It may be noted that, even though E is a vector space, the metric is not derived from a norm since, if $\lambda \in \mathbb{R}$, one may have $\mathrm{d}(\lambda g, \lambda h) \ne |\lambda|\mathrm{d}(g, h)$.

(v) Let E be the set of all measurable functions $f: I \to \mathbb{R}$, where $I = [a, b]$ is an interval of $\mathbb{R}$, and define

$$\mathrm{d}(g, h) = \int_a^b |g(x) - h(x)|(1 + |g(x) - h(x)|)^{-1} dx.$$

In order to obtain **(D1)**, we identify functions which take the same value at all points of I, except for a set of measure zero.

Convergence with respect to this metric coincides with *convergence in measure*, which plays a role in the theory of probability.

(vi) Let $E = \mathbb{F}_2^\infty$ be the set of all infinite sequences $a = (\alpha_1, \alpha_2, \ldots)$, where $\alpha_j = 0$ or 1 for every j, and define $\mathrm{d}(a, a) = 0$, $\mathrm{d}(a, b) = 2^{-k}$ if $a \ne b$, where $b = (\beta_1, \beta_2, \ldots)$ and k is the least positive integer such that $\alpha_k \ne \beta_k$.

Here the triangle inequality holds in the stronger form

$$\mathrm{d}(a, b) \leq \max[\mathrm{d}(a, c), \mathrm{d}(c, b)].$$

This metric space plays a basic role in the theory of *dynamical systems*.

(vii) A connected *graph* can be given the structure of a metric space by defining the distance between two vertices to be the number of edges on the shortest path joining them.

Let E be an arbitrary metric space and $\{a_n\}$ a sequence of elements of E. The sequence $\{a_n\}$ is said to *converge*, with *limit* $a \in E$, if

$$\mathrm{d}(a_n, a) \to 0 \quad \text{as } n \to \infty,$$

i.e. if for each real $\varepsilon > 0$ there is a corresponding positive integer $N = N(\varepsilon)$ such that $\mathrm{d}(a_n, a) < \varepsilon$ for every $n \geq N$.

The limit a is uniquely determined, since if also $\mathrm{d}(a_n, a') \to 0$, then

$$\mathrm{d}(a, a') \leq \mathrm{d}(a_n, a) + \mathrm{d}(a_n, a'),$$

and the right side can be made arbitrarily small by taking n sufficiently large. We write

$$\lim_{n\to\infty} a_n = a,$$

or $a_n \to a$ as $n \to \infty$. If the sequence $\{a_n\}$ has limit a, then so also does any (infinite) subsequence.

If $a_n \to a$ and $b_n \to b$, then $\mathrm{d}(a_n, b_n) \to \mathrm{d}(a, b)$, as one sees by taking $a' = a_n$ and $b' = b_n$ in $(*)$.

The sequence $\{a_n\}$ is said to be a *fundamental sequence*, or 'Cauchy sequence', if for each real $\varepsilon > 0$ there is a corresponding positive integer $N = N(\varepsilon)$ such that $\mathrm{d}(a_m, a_n) < \varepsilon$ for all $m, n \geq N$.

If $\{a_n\}$ and $\{b_n\}$ are fundamental sequences then, by $(*)$, the sequence $\{\mathrm{d}(a_n, b_n)\}$ of real numbers is a fundamental sequence, and therefore convergent.

A set $S \subseteq E$ is said to be *bounded* if the set of all real numbers $\mathrm{d}(a, b)$ with $a, b \in S$ is a bounded subset of $\mathbb{R}$.

Any fundamental sequence $\{a_n\}$ is bounded, since if

$$\mathrm{d}(a_m, a_n) < 1 \quad \text{for all } m, n \geq N,$$

then

$$\mathrm{d}(a_m, a_n) < 1 + \delta \quad \text{for all } m, n \in \mathbb{N},$$

where $\delta = \max_{1 \leq j < k \leq N} \mathrm{d}(a_j, a_k)$.

Furthermore, any convergent sequence $\{a_n\}$ is a fundamental sequence, as one sees by taking $a = \lim_{n\to\infty} a_n$ in the inequality

$$\mathrm{d}(a_m, a_n) \leq \mathrm{d}(a_m, a) + \mathrm{d}(a_n, a).$$

A metric space is said to be *complete* if, conversely, every fundamental sequence is convergent.

By generalizing the Méray–Cantor method of extending the rational numbers to the real numbers, Hausdorff (1913) showed that any metric space can be embedded in a complete metric space. To state his result precisely, we introduce some definitions.

A subset F of a metric space E is said to be *dense* in E if, for each $a \in E$ and each real $\varepsilon > 0$, there exists some $b \in F$ such that $\mathrm{d}(a, b) < \varepsilon$.

A map σ from one metric space E to another metric space E' is necessarily injective if it is distance-preserving, i.e. if

$$\mathrm{d}'(\sigma(a), \sigma(b)) = \mathrm{d}(a, b) \quad \text{for all } a, b \in E.$$

If the map σ is also surjective, then it is said to be an *isometry* and the metric spaces E and E' are said to be *isometric*.

A metric space $\bar{E}$ is said to be a *completion* of a metric space E if $\bar{E}$ is complete and E is isometric to a dense subset of $\bar{E}$. It is easily seen that any two completions of a given metric space are isometric.

Hausdorff's result says that *any metric space E has a completion $\bar{E}$*. We sketch the proof. Define two fundamental sequences $\{a_n\}$ and $\{a'_n\}$ in E to be equivalent if

$$\lim_{n\to\infty} \mathrm{d}(a_n, a'_n) = 0.$$

It is easily shown that this is indeed an equivalence relation. Moreover, if the fundamental sequences $\{a_n\}, \{b_n\}$ are equivalent to the fundamental sequences $\{a'_n\}, \{b'_n\}$ respectively, then

$$\lim_{n\to\infty} \mathrm{d}(a_n, b_n) = \lim_{n\to\infty} \mathrm{d}(a'_n, b'_n).$$

We can give the set $\bar{E}$ of all equivalence classes of fundamental sequences the structure of a metric space by defining

$$\bar{d}(\{a_n\}, \{b_n\}) = \lim_{n\to\infty} \mathrm{d}(a_n, b_n).$$

For each $a \in E$, let $\bar{a}$ be the equivalence class in $\bar{E}$ which contains the fundamental sequence $\{a_n\}$ such that $a_n = a$ for every n. Since

$$\bar{d}(\bar{a}, \bar{b}) = \mathrm{d}(a, b) \quad \text{for all } a, b \in E,$$

E is isometric to the set $E' = \{\bar{a} : a \in E\}$. It is not difficult to show that E' is dense in $\bar{E}$ and that $\bar{E}$ is complete.

Which of the previous examples of metric spaces are complete? In example (i), the completeness of $\mathbb{R}^n$ with respect to the first definition of distance follows directly from the completeness of $\mathbb{R}$. It is also complete with respect to the two alternative definitions of distance, since a sequence which converges with respect to one of the three metrics also converges with respect to the other two. Indeed it is easily shown that, for every $a \in \mathbb{R}^n$,

$$|a| \le |a|_2 \le |a|_1$$

and

$$|a|_1 \le n^{1/2}|a|_2, \quad |a|_2 \le n^{1/2}|a|.$$

In example (ii), the completeness of $\mathbb{F}_2^n$ is trivial, since any fundamental sequence is ultimately constant.

In example (iii), the completeness of $\mathscr{C}(I)$ with respect to the first definition of distance follows from the completeness of $\mathbb{R}$ and the fact that the limit of a uniformly convergent sequence of continuous functions is again a continuous function.

However, $\mathscr{C}(I)$ is not complete with respect to either of the two alternative definitions of distance. It is possible also for a sequence to converge with respect to the two alternative definitions of distance, but not with respect to the first definition. Similarly, a sequence may converge in the first alternative metric, but not even be a fundamental sequence in the second.

The completions of the metric space $\mathscr{C}(I)$ with respect to the two alternative metrics may actually be identified with spaces of functions. The completion for the first alternative metric is the set $L(I)$ of all *Lebesgue measurable* functions $f: I \to \mathbb{R}$ such that

$$\int_a^b |f(x)|dx < \infty,$$

functions which take the same value at all points of I, except for a set of measure zero, being identified. The completion $L^2(I)$ for the second alternative metric is obtained by replacing $\int_a^b |f(x)|dx$ by $\int_a^b |f(x)|^2 dx$ in this statement.

It may be shown that the metric spaces of examples (iv)–(vi) are all complete. In example (vi), the strong triangle inequality implies that $\{a_n\}$ is a fundamental sequence if (and only if) $\mathrm{d}(a_{n+1}, a_n) \to 0$ as $n \to \infty$.

Let E be an arbitrary metric space and $f: E \to E$ a map of E into itself. A point $\bar{x} \in E$ is said to be a *fixed point* of f if $f(\bar{x}) = \bar{x}$. A useful property of complete metric spaces is the following *contraction principle*, which was first established in the present generality by Banach (1922), but was previously known in more concrete situations.

Proposition 26 *Let E be a complete metric space and let $f: E \to E$ be a map of E into itself. If there exists a real number θ, with $0 < \theta < 1$, such that*

$$\mathrm{d}(f(x'), f(x'')) \leq \theta \mathrm{d}(x', x'') \quad \textit{for all } x', x'' \in E,$$

then the map f has a unique fixed point $\bar{x} \in E$.

Proof It is clear that there is at most one fixed point, since $0 \leq \mathrm{d}(x', x'') \leq \theta \mathrm{d}(x', x'')$ implies $x' = x''$. To prove that a fixed point exists we use the *method of successive approximations*.

Choose any $x_0 \in E$ and define the sequence $\{x_n\}$ recursively by

$$x_n = f(x_{n-1}) \quad (n \geq 1).$$

For any $k \geq 1$ we have

$$\mathrm{d}(x_{k+1}, x_k) = \mathrm{d}(f(x_k), f(x_{k-1})) \leq \theta \mathrm{d}(x_k, x_{k-1}).$$

Applying this k times, we obtain

$$\mathrm{d}(x_{k+1}, x_k) \leq \theta^k \mathrm{d}(x_1, x_0).$$

Consequently, if $n > m \geq 0$,

$$\begin{aligned} \mathrm{d}(x_n, x_m) &\leq \mathrm{d}(x_n, x_{n-1}) + \mathrm{d}(x_{n-1}, x_{n-2}) + \cdots + \mathrm{d}(x_{m+1}, x_m) \\ &\leq (\theta^{n-1} + \theta^{n-2} + \cdots + \theta^m)\mathrm{d}(x_1, x_0) \\ &\leq \theta^m (1-\theta)^{-1} \mathrm{d}(x_1, x_0), \end{aligned}$$

since $0 < \theta < 1$. It follows that $\{x_n\}$ is a fundamental sequence and so a convergent sequence, since E is complete. If $\bar{x} = \lim_{n\to\infty} x_n$, then

$$\begin{aligned} \mathrm{d}(f(\bar{x}), \bar{x}) &\leq \mathrm{d}(f(\bar{x}), x_{n+1}) + \mathrm{d}(x_{n+1}, \bar{x}) \\ &\leq \theta \mathrm{d}(\bar{x}, x_n) + \mathrm{d}(\bar{x}, x_{n+1}). \end{aligned}$$

Since the right side can be made less than any given positive real number by taking n large enough, we must have $f(\bar{x}) = \bar{x}$. The proof shows also that, for any $m \geq 0$,

$$\mathrm{d}(\bar{x}, x_m) \leq \theta^m (1-\theta)^{-1} \mathrm{d}(x_1, x_0). \qquad \square$$

The contraction principle is surprisingly powerful, considering the simplicity of its proof. We give two significant applications: an inverse function theorem and an existence theorem for ordinary differential equations. In both cases we will use the notion of differentiability for functions of several real variables. The unambitious reader may simply take $n = 1$ in the following discussion (so that 'invertible' means 'nonzero'). Functions of several variables are important, however, and it is remarkable that the proper definition of differentiability in this case was first given by Stolz (1887).

A map $\varphi\colon U \to \mathbb{R}^m$, where $U \subseteq \mathbb{R}^n$ is a *neighbourhood* of $x_0 \in \mathbb{R}^n$ (i.e., U contains some open ball $\{x \in \mathbb{R}^n : |x - x_0| < \rho\}$), is said to be *differentiable* at x_0 if there exists a linear map $A\colon \mathbb{R}^n \to \mathbb{R}^m$ such that

$$|\varphi(x) - \varphi(x_0) - A(x - x_0)|/|x - x_0| \to 0 \quad \text{as } |x - x_0| \to 0.$$

(The inequalities between the various norms show that it is immaterial which norm is used.) The linear map A, which is then uniquely determined, is called the *derivative* of φ at x_0 and will be denoted by $\varphi'(x_0)$.

This definition is a natural generalization of the usual definition when $m = n = 1$, since it says that the difference $\varphi(x_0 + h) - \varphi(x_0)$ admits the linear approximation Ah for $|h| \to 0$.

Evidently, if φ_1 and φ_2 are differentiable at x_0, then so also is $\varphi = \varphi_1 + \varphi_2$ and

$$\varphi'(x_0) = \varphi_1'(x_0) + \varphi_2'(x_0).$$

It also follows directly from the definition that derivatives satisfy the *chain rule*: If $\varphi\colon U \to \mathbb{R}^m$, where U is a neighbourhood of $x_0 \in \mathbb{R}^n$, is differentiable at x_0, and if $\psi\colon V \to \mathbb{R}^l$, where V is a neighbourhood of $y_0 = \varphi(x_0) \in \mathbb{R}^m$, is differentiable at y_0, then the composite map $\chi = \psi \circ \varphi\colon U \to \mathbb{R}^l$ is differentiable at x_0 and

$$\chi'(x_0) = \psi'(y_0)\varphi'(x_0),$$

the right side being the composite linear map.

We will also use the notion of norm of a linear map. If $A: \mathbb{R}^n \to \mathbb{R}^m$ is a linear map, its *norm* $|A|$ is defined by

$$|A| = \sup_{|x| \le 1} |Ax|.$$

Evidently

$$|A_1 + A_2| \le |A_1| + |A_2|.$$

Furthermore, if $B: \mathbb{R}^m \to \mathbb{R}^l$ is another linear map, then

$$|BA| \le |B||A|.$$

Hence, if $m = n$ and $|A| < 1$, then the linear map $I - A$ is invertible, its inverse being given by the geometric series

$$(I - A)^{-1} = I + A + A^2 + \cdots .$$

It follows that for any invertible linear map $A: \mathbb{R}^n \to \mathbb{R}^n$, if $B: \mathbb{R}^n \to \mathbb{R}^n$ is a linear map such that $|B - A| < |A^{-1}|^{-1}$, then B is also invertible and $|B^{-1} - A^{-1}| \to 0$ as $|B - A| \to 0$.

If $\varphi: U \to \mathbb{R}^m$ is differentiable at $x_0 \in \mathbb{R}^n$, then it is also continuous at x_0, since

$$|\varphi(x) - \varphi(x_0)| \le |\varphi(x) - \varphi(x_0) - \varphi'(x_0)(x - x_0)| + |\varphi'(x_0)||x - x_0|.$$

We say that φ is *continuously differentiable* in U if it is differentiable at each point of U and if the derivative $\varphi'(x)$ is a continuous function of x in U. The *inverse function theorem* says:

Proposition 27 *Let U_0 be a neighbourhood of $x_0 \in \mathbb{R}^n$ and let $\varphi: U_0 \to \mathbb{R}^n$ be a continuously differentiable map for which $\varphi'(x_0)$ is invertible.*

Then, for some $\delta > 0$, the ball $U = \{x \in \mathbb{R}^n : |x - x_0| < \delta\}$ is contained in U_0 and

(i) *the restriction of φ to U is injective;*
(ii) *$V := \varphi(U)$ is open, i.e. if $\eta \in V$, then V contains all $y \in \mathbb{R}^n$ near η;*
(iii) *the inverse map $\psi: V \to U$ is also continuously differentiable and, if $y = \varphi(x)$, then $\psi'(y)$ is the inverse of $\varphi'(x)$.*

Proof To simplify notation, assume $x_0 = \varphi(x_0) = 0$ and write $A = \varphi'(0)$. For any $y \in \mathbb{R}^n$, put

$$f_y(x) = x + A^{-1}[y - \varphi(x)].$$

Evidently x is a fixed point of f_y if and only if $\varphi(x) = y$. The map f_y is also continuously differentiable and

$$f_y'(x) = I - A^{-1}\varphi'(x) = A^{-1}[A - \varphi'(x)].$$

Since $\varphi'(x)$ is continuous, we can choose $\delta > 0$ so that the ball $U = \{x \in \mathbb{R}^n : |x| < \delta\}$ is contained in U_0 and

$$|f_y'(x)| \le 1/2 \quad \text{for } x \in U.$$

If $x_1, x_2 \in U$, then

$$|f_y(x_2) - f_y(x_1)| = \left| \int_0^1 f'((1-t)x_1 + tx_2)(x_2 - x_1)dt \right| \\ \leq |x_2 - x_1|/2.$$

It follows that f_y has at most one fixed point in U. Since this holds for arbitrary $y \in \mathbb{R}^n$, the restriction of φ to U is injective.

Suppose next that $\eta = \varphi(\xi)$ for some $\xi \in U$. We wish to show that, if y is near η, the map f_y has a fixed point near ξ.

Choose $r = r(\xi) > 0$ so that the closed ball $B_r = \{x \in \mathbb{R}^n : |x - \xi| \leq r\}$ is contained in U, and fix $y \in \mathbb{R}^n$ so that $|y - \eta| < r/2|A^{-1}|$. Then

$$|f_y(\xi) - \xi| = |A^{-1}(y - \eta)| \\ \leq |A^{-1}||y - \eta| < r/2.$$

Hence if $|x - \xi| \leq r$, then

$$|f_y(x) - \xi| \leq |f_y(x) - f_y(\xi)| + |f_y(\xi) - \xi| \\ \leq |x - \xi|/2 + r/2 \leq r.$$

Thus $f_y(B_r) \subseteq B_r$. Also, if $x_1, x_2 \in B_r$, then

$$|f_y(x_2) - f_y(x_1)| \leq |x_2 - x_1|/2.$$

But B_r is a complete metric space, with the same metric as $\mathbb{R}^n$, since it is a closed subset (if $x_n \in B_r$ and $x_n \to x$ in $\mathbb{R}^n$, then also $x \in B_r$). Consequently, by the contraction principle (Proposition 26), f_y has a fixed point $x \in B_r$. Then $\varphi(x) = y$, which proves (ii).

Suppose now that $y, \eta \in V$. Then $y = \varphi(x)$, $\eta = \varphi(\xi)$ for unique $x, \xi \in U$. Since

$$|f_y(x) - f_y(\xi)| \leq |x - \xi|/2$$

and

$$f_y(x) - f_y(\xi) = x - \xi - A^{-1}(y - \eta),$$

we have

$$|A^{-1}(y - \eta)| \geq |x - \xi|/2.$$

Thus

$$|x - \xi| \leq 2|A^{-1}||y - \eta|.$$

If $F = \varphi'(\xi)$ and $G = F^{-1}$, then

$$\psi(y) - \psi(\eta) - G(y - \eta) = x - \xi - G(y - \eta) \\ = -G[\varphi(x) - \varphi(\xi) - F(x - \xi)].$$

Hence

$$|\psi(y) - \psi(\eta) - G(y-\eta)|/|y-\eta| \leq 2|A^{-1}||G||\varphi(x) - \varphi(\xi) - F(x-\xi)|/|x-\xi|.$$

If $|y-\eta| \to 0$, then $|x-\xi| \to 0$ and the right side tends to 0. Consequently ψ is differentiable at η and $\psi'(\eta) = G = F^{-1}$.

Thus ψ is differentiable in U and, *a fortiori*, continuous. In fact ψ is continuously differentiable, since F is a continuous function of ξ (by hypothesis), since $\xi = \psi(\eta)$ is a continuous function of η, and since F^{-1} is a continuous function of F. □

To bring out the meaning of Proposition 27 we add some remarks:

(i) The invertibility of $\varphi'(x_0)$ is necessary for the existence of a differentiable inverse map, but not for the existence of a continuous inverse map. For example, the continuously differentiable map $\varphi: \mathbb{R} \to \mathbb{R}$ defined by $\varphi(x) = x^3$ is bijective and has the continuous inverse $\psi(y) = y^{1/3}$, although $\varphi'(0) = 0$.

(ii) The hypothesis that φ is *continuously* differentiable cannot be totally dispensed with. For example, the map $\varphi: \mathbb{R} \to \mathbb{R}$ defined by

$$\varphi(x) = x + x^2 \sin(1/x) \quad \text{if } x \neq 0, \varphi(0) = 0,$$

is everywhere differentiable and $\varphi'(0) \neq 0$, but φ is not injective in any neighbourhood of 0.

(iii) The inverse map may not be defined throughout U_0. For example, the map $\varphi: \mathbb{R}^2 \to \mathbb{R}^2$ defined by

$$\varphi_1(x_1, x_2) = x_1^2 - x_2^2, \quad \varphi_2(x_1, x_2) = 2x_1x_2,$$

is everywhere continuously differentiable and has an invertible derivative at every point except the origin. Thus the hypotheses of Proposition 27 are satisfied in any connected open set $U_0 \subseteq \mathbb{R}^2$ which does not contain the origin, and yet $\varphi(1, 1) = \varphi(-1, -1)$.

It was first shown by Cauchy (c. 1844) that, under quite general conditions, an ordinary differential equation has local solutions. The method of successive approximations (i.e., the contraction principle) was used for this purpose by Picard (1890):

Proposition 28 *Let $t_0 \in \mathbb{R}, \xi_0 \in \mathbb{R}^n$ and let U be a neighbourhood of (t_0, ξ_0) in $\mathbb{R} \times \mathbb{R}^n$. If $\varphi: U \to \mathbb{R}^n$ is a continuous map with a derivative φ' with respect to x that is continuous in U, then the differential equation*

$$dx/dt = \varphi(t, x) \tag{1}$$

has a unique solution $x(t)$ which satisfies the initial condition

$$x(t_0) = \xi_0 \tag{2}$$

and is defined in some interval $|t - t_0| \leq \delta$, where $\delta > 0$.

Proof If $x(t)$ is a solution of the differential equation (1) which satisfies the initial condition (2), then by integration we get

$$x(t_0) = \xi_0 + \int_{t_0}^{t} \varphi[\tau, x(\tau)]d\tau.$$

Conversely, if a *continuous* function $x(t)$ satisfies this relation then, since φ is continuous, $x(t)$ is actually differentiable and is a solution of (1) that satisfies (2). Hence we need only show that the map $\mathscr{F}$ defined by

$$(\mathscr{F}x)(t) = \xi_0 + \int_{t_0}^{t} \varphi[\tau, x(\tau)]d\tau$$

has a unique fixed point in the space of continuous functions.

There exist positive constants M, L such that

$$|\varphi(t, \xi)| \leq M, \quad |\varphi'(t, \xi)| \leq L$$

for all (t, ξ) in a neighbourhood of (t_0, ξ_0), which we may take to be U. If $(t, \xi_1) \in U$ and $(t, \xi_2) \in U$, then

$$\begin{aligned} |\varphi(t, \xi_2) - \varphi(t, \xi_1)| &= \left| \int_0^1 \varphi'(t, (1-u)\xi_1 + u\xi_2)(\xi_2 - \xi_1)du \right| \\ &\leq L|\xi_2 - \xi_1|. \end{aligned}$$

Choose $\delta > 0$ so that the box $|t - t_0| \leq \delta, |\xi - \xi_0| \leq M\delta$ is contained in U and also $L\delta < 1$. Take $I = [t_0 - \delta, t_0 + \delta]$ and let $\mathscr{C}(I)$ be the complete metric space of all continuous functions $x : I \to \mathbb{R}^n$ with the distance function

$$\mathrm{d}(x_1, x_2) = \sup_{t \in I} |x_1(t) - x_2(t)|.$$

The constant function $x_0(t) = \xi_0$ is certainly in $\mathscr{C}(I)$. Let E be the subset of all $x \in \mathscr{C}(I)$ such that $x(t_0) = \xi_0$ and $\mathrm{d}(x, x_0) \leq M\delta$. Evidently if $x_n \in E$ and $x_n \to x$ in $\mathscr{C}(I)$, then $x \in E$. Hence E is also a complete metric space with the same metric. Moreover $\mathscr{F}(E) \subseteq E$, since if $x \in E$ then $(\mathscr{F}x)(t_0) = \xi_0$ and, for all $t \in I$,

$$|(\mathscr{F}x)(t) - \xi_0| = \left| \int_{t_0}^{t} \varphi[\tau, x(\tau)]d\tau \right| \leq M\delta.$$

Furthermore, if $x_1, x_2 \in E$, then $\mathrm{d}(\mathscr{F}x_1, \mathscr{F}x_2) \leq L\delta \mathrm{d}(x_1, x_2)$, since for all $t \in I$,

$$\begin{aligned} |(\mathscr{F}x_1)(t) - (\mathscr{F}x_2)(t)| &= \left| \int_{t_0}^{t} \{\varphi[\tau, x_1(\tau)] - \varphi[\tau, x_2(\tau)]\}d\tau \right| \\ &\leq L\delta\, \mathrm{d}(x_1, x_2). \end{aligned}$$

Since $L\delta < 1$, the result now follows from Proposition 26. □

Proposition 28 only guarantees the local existence of solutions, but this is in the nature of things. For example, if $n = 1$, the unique solution of the differential equation

$$dx/dt = x^2$$

such that $x(t_0) = \xi_0 > 0$ is given by

$$x(t) = \{1 - (t - t_0)\xi_0\}^{-1}\xi_0.$$

Thus the solution is defined only for $t < t_0 + \xi_0^{-1}$, even though the differential equation itself has exemplary behaviour everywhere.

To illustrate Proposition 28, take $n = 1$ and let $E(t)$ be the solution of the (linear) differential equation

$$dx/dt = x \tag{3}$$

which satisfies the initial condition $E(0) = 1$. Then $E(t)$ is defined for $|t| < R$, for some $R > 0$. If $|\tau| < R/2$ and $x_1(t) = E(t + \tau)$, then $x_1(t)$ is the solution of the differential equation (3) which satisfies the initial condition $x_1(0) = E(\tau)$. But $x_2(t) = E(\tau)E(t)$ satisfies the same differential equation and the same initial condition. Hence we must have $x_1(t) = x_2(t)$ for $|t| < R/2$, i.e.

$$E(t + \tau) = E(t)E(\tau). \tag{4}$$

In particular,

$$E(t)E(-t) = 1, \quad E(2t) = E(t)^2.$$

The last relation may be used to extend the definition of $E(t)$, so that it is continuously differentiable and a solution of (3) also for $|t| < 2R$. It follows that the solution $E(t)$ is defined for all $t \in \mathbb{R}$ and satisfies the *addition theorem* (4) for all $t, \tau \in \mathbb{R}$.

It is instructive to carry through the method of successive approximations explicitly in this case. If we take $x_0(t)$ to be the constant 1, then

$$x_1(t) = 1 + \int_0^t x_0(\tau)d\tau = 1 + t,$$
$$x_2(t) = 1 + \int_0^t x_1(\tau)d\tau = 1 + t + t^2/2,$$
$$\cdots .$$

By induction we obtain, for every $n \geq 1$,

$$x_n(t) = 1 + t + t^2/2! + \cdots + t^n/n!.$$

Since $x_n(t) \to E(t)$ as $n \to \infty$, we obtain for the solution $E(t)$ the infinite series representation

$$E(t) = 1 + t + t^2/2! + t^3/3! + \cdots ,$$

valid actually for every $t \in \mathbb{R}$. In particular,

$$e := E(1) = 1 + 1 + 1/2! + 1/3! + \cdots = 2.7182818\ldots .$$

Of course $E(t) = e^t$ is the *exponential function*. We will now adopt the usual notation, but we remark that the definition of e^t as a solution of a differential equation provides a meaning for irrational t, as well as a simple proof of both the addition theorem and the exponential series.

The power series for e^t shows that

$$e^t > 1 + t > 1 \quad \text{for every } t > 0.$$

Since $e^{-t} = (e^t)^{-1}$, it follows that $0 < e^t < 1$ for every $t < 0$. Thus $e^t > 0$ for all $t \in \mathbb{R}$. Hence, by (3), e^t is a strictly increasing function. But $e^t \to +\infty$ as $t \to +\infty$ and $e^t \to 0$ as $t \to -\infty$. Consequently, since it is certainly continuous, the exponential function maps the real line $\mathbb{R}$ bijectively onto the positive half-line $\mathbb{R}_+ = \{x \in R : x > 0\}$. For any $x > 0$, the unique $t \in \mathbb{R}$ such that $e^t = x$ is denoted by $\ln x$ (the *natural logarithm* of x) or simply $\log x$.

5 Complex Numbers

By extending the rational numbers to the real numbers, we ensured that every positive number had a square root. By further extending the real numbers to the complex numbers, we will now ensure that all numbers have square roots.

The first use of complex numbers, by Cardano (1545), may have had its origin in the solution of cubic, rather than quadratic, equations. The cubic polynomial

$$f(x) = x^3 - 3px - 2q$$

has three real roots if $d := q^2 - p^3 < 0$ since then, for large $X > 0$,

$$f(-X) < 0, \quad f(-p^{1/2}) > 0, \quad f(p^{1/2}) < 0, \quad f(X) > 0.$$

Cardano's formula for the three roots,

$$f(x) = \sqrt[3]{(q + \sqrt{d})} + \sqrt[3]{(q - \sqrt{d})},$$

gives real values, even though d is negative, because the two summands are conjugate complex numbers. This was explicitly stated by Bombelli (1572). It is a curious fact, first proved by Hölder (1891), that if a cubic equation has three distinct real roots, then it is impossible to represent these roots solely by real radicals.

Intuitively, complex numbers are expressions of the form $a + ib$, where a and b are real numbers and $i^2 = -1$. But what is i? Hamilton (1835) defined complex numbers as ordered pairs of real numbers, with appropriate rules for addition and multiplication. Although this approach is similar to that already used in this chapter, and actually was its first appearance, we now choose a different method.

We define a *complex number* to be a 2×2 matrix of the form

$$A = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}.$$

where a and b are real numbers. The set of all complex numbers is customarily denoted by $\mathbb{C}$. We may define addition and multiplication in $\mathbb{C}$ to be matrix addition and multiplication, since $\mathbb{C}$ is closed under these operations: if

$$B = \begin{pmatrix} c & d \\ -d & c \end{pmatrix},$$

then

$$A + B = \begin{pmatrix} a+c & b+d \\ -(b+d) & a+c \end{pmatrix}, \quad AB = \begin{pmatrix} ac-bd & ad+bc \\ -(ad+bc) & ac-bd \end{pmatrix}.$$

Furthermore $\mathbb{C}$ contains

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

and $A \in \mathbb{C}$ implies $-A \in \mathbb{C}$.

It follows from the properties of matrix addition and multiplication that addition and multiplication of complex numbers have the properties **(A2)–(A5)**, **(M2)–(M4)** and **(AM1)–(AM2)**, with 0 and 1 as identity elements for addition and multiplication respectively. The property **(M5)** also holds, since if a and b are not both zero, and if

$$a' = a/(a^2 + b^2), \quad b' = -b/(a^2 + b^2),$$

then

$$A^{-1} = \begin{pmatrix} a' & b' \\ -b' & a' \end{pmatrix}$$

is a multiplicative inverse of A. Thus $\mathbb{C}$ satisfies the axioms for a *field*.

The set $\mathbb{C}$ also contains the matrix

$$i = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

for which $i^2 = -1$, and any $A \in \mathbb{C}$ can be represented in the form

$$A = a1 + bi,$$

where $a, b \in \mathbb{R}$. The multiples $a1$, where $a \in \mathbb{R}$, form a subfield of $\mathbb{C}$ isomorphic to the real field $\mathbb{R}$. By identifying the real number a with the complex number $a1$, we may regard $\mathbb{R}$ itself as contained in $\mathbb{C}$.

Thus we will now stop using matrices and use only the fact that $\mathbb{C}$ is a field containing $\mathbb{R}$ such that every $z \in \mathbb{C}$ can be represented in the form

$$z = x + iy,$$

where $x, y \in \mathbb{R}$ and $i \in \mathbb{C}$ satisfies $i^2 = -1$. The representation is necessarily unique, since $i \notin \mathbb{R}$. We call x and y the *real* and *imaginary parts* of z and denote them by $\mathscr{R}z$,

and $\mathscr{I}z$ respectively. Complex numbers of the form iy, where $y \in \mathbb{R}$, are said to be *pure imaginary*.

It is worth noting that $\mathbb{C}$ cannot be given the structure of an *ordered* field, since in an ordered field any nonzero square is positive, whereas $i^2 + 1^2 = (-1) + 1 = 0$.

It is often suggestive to regard complex numbers as points of a plane, the complex number $z = x + iy$ being the point with coordinates (x, y) in some chosen system of rectangular coordinates.

The *complex conjugate* of the complex number $z = x + iy$, where $x, y \in \mathbb{R}$, is the complex number $\bar{z} = x - iy$. In the geometrical representation of complex numbers, $\bar{z}$ is the reflection of z in the x-axis. From the definition we at once obtain

$$\mathscr{R}z = (z + \bar{z})/2, \quad \mathscr{I}z = (z - \bar{z})/2i.$$

It is easily seen also that

$$\overline{z + w} = \bar{z} + \bar{w}, \quad \overline{zw} = \bar{z}\bar{w}, \quad \bar{\bar{z}} = z.$$

Moreover, $\bar{z} = z$ if and only if $z \in \mathbb{R}$. Thus the map $z \to \bar{z}$ is an 'involutory automorphism' of the field $\mathbb{C}$, with the subfield $\mathbb{R}$ as its set of fixed points. It follows that $\overline{-z} = -\bar{z}$.

If $z = x + iy$, where $x, y \in \mathbb{R}$, then

$$z\bar{z} = (x + iy)(x - iy) = x^2 + y^2.$$

Hence $z\bar{z}$ is a positive real number for any nonzero $z \in \mathbb{C}$. The *absolute value* $|z|$ of the complex number z is defined by

$$|0| = 0, \quad |z| = \sqrt{(z\bar{z})} \text{ if } z \neq 0,$$

(with the positive value for the square root). This agrees with the definition in §3 if $z = x$ is a positive real number.

It follows at once from the definition that $|\bar{z}| = |z|$ for every $z \in \mathbb{C}$, and $z^{-1} = \bar{z}/|z|^2$ if $z \neq 0$.

Absolute values have the following properties: *for all* $z, w \in \mathbb{C}$,

(i) $|0| = 0, |z| > 0$ *if* $z \neq 0$;
(ii) $|zw| = |z||w|$;
(iii) $|z + w| \leq |z| + |w|$.

The first property follows at once from the definition. To prove (ii), observe that both sides are non-negative and that

$$|zw|^2 = zw\overline{zw} = zw\bar{z}\bar{w} = z\bar{z}w\bar{w} = |z|^2|w|^2.$$

To prove (iii), we first evaluate $|z + w|^2$:

$$|z + w|^2 = (z + w)(\bar{z} + \bar{w}) = z\bar{z} + (z\bar{w} + w\bar{z}) + w\bar{w} = |z|^2 + 2\mathscr{R}(z\bar{w}) + |w|^2.$$

Since $\mathscr{R}(z\bar{w}) \leq |z\bar{w}| = |z||w|$, this yields

$$|z+w|^2 \le |z|^2 + 2|z||w| + |w|^2 = (|z|+|w|)^2,$$

and (iii) follows by taking square roots.

Several other properties are consequences of these three, although they may also be verified directly. By taking $z = w = 1$ in (ii) and using (i), we obtain $|1| = 1$. By taking $z = w = -1$ in (ii) and using (i), we now obtain $|-1| = 1$. Taking $w = -1$ and $w = z^{-1}$ in (ii), we further obtain

$$|-z| = |z|, \quad |z^{-1}| = |z|^{-1} \quad \text{if } z \neq 0.$$

Again, by replacing z by $z - w$ in (iii), we obtain

$$||z| - |w|| \le |z - w|.$$

This shows that $|z|$ is a continuous function of z. In fact $\mathbb{C}$ is a metric space, with the metric $\mathrm{d}(z, w) = |z - w|$. By considering real and imaginary parts separately, one verifies that this metric space is complete, i.e. every fundamental sequence is convergent, and that the Bolzano–Weierstrass property continues to hold, i.e. any bounded sequence of complex numbers has a convergent subsequence.

It will now be shown that any complex number has a square root. If $w = u + iv$ and $z = x + iy$, then $z^2 = w$ is equivalent to

$$x^2 - y^2 = u, \quad 2xy = v.$$

Since

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

these equations imply

$$x^2 + y^2 = \sqrt{(u^2 + v^2)}.$$

Hence

$$x^2 = \{u + \sqrt{(u^2 + v^2)}\}/2.$$

Since the right side is positive if $v \neq 0$, x is then uniquely determined apart from sign and $y = v/2x$ is uniquely determined by x. If $v = 0$, then $x = \pm\sqrt{u}$ and $y = 0$ when $u > 0$; $x = 0$ and $y = \pm\sqrt{(-u)}$ when $u < 0$, and $x = y = 0$ when $u = 0$.

It follows that any quadratic polynomial

$$q(z) = az^2 + bz + c,$$

where $a, b, c \in \mathbb{C}$ and $a \neq 0$, has two complex roots, given by the well-known formula

$$z = \{-b \pm \sqrt{(b^2 - 4ac)}\}/2a.$$

However, much more is true. The so-called *fundamental theorem of algebra* asserts that any polynomial

$$f(z) = a_0 z^n + a_1 z^{n-1} + \cdots + a_n,$$

where $a_0, a_1, \ldots, a_n \in \mathbb{C}$, $n \geq 1$ and $a_0 \neq 0$, has a complex root. Thus by adjoining to the real field $\mathbb{R}$ a root of the polynomial $z^2 + 1$ we ensure that every non-constant polynomial has a root. Today the fundamental theorem of algebra is considered to belong to analysis, rather than to algebra. It is useful to retain the name, however, as a reminder that our own pronouncements may seem equally quaint in the future.

Our proof of the theorem will use the fact that any polynomial is differentiable, since sums and products of differentiable functions are again differentiable, and hence also continuous. We first prove

Proposition 29 *Let $G \subseteq \mathbb{C}$ be an open set and E a proper subset (possibly empty) of G such that each point of G has a neighbourhood containing at most one point of E. If $f: G \to \mathbb{C}$ is a continuous map which at every point of $G \backslash E$ is differentiable and has a nonzero derivative, then $f(G)$ is an open subset of $\mathbb{C}$.*

Proof Evidently $G \backslash E$ is an open set. We show first that $f(G \backslash E)$ is also an open set. Let $\zeta \in G \backslash E$. Then f is differentiable at ζ and $\rho = |f'(\zeta)| > 0$. We can choose $\delta > 0$ so that the closed disc $B = \{z \in \mathbb{C}: |z - \zeta| \leq \delta\}$ contains no point of E, is contained in G and

$$|f(z) - f(\zeta)| \geq \rho |z - \zeta|/2 \quad \text{for every } z \in B.$$

In particular, if $S = \{z \in \mathbb{C}: |z - \zeta| = \delta\}$ is the boundary of B, then

$$|f(z) - f(\zeta)| \geq \rho\delta/2 \quad \text{for every } z \in S.$$

Choose $w \in \mathbb{C}$ so that $|w - f(\zeta)| < \rho\delta/4$ and consider the minimum in the compact set B of the continuous real-valued function $\phi(z) = |f(z) - w|$. On the boundary S we have

$$\phi(z) \geq |f(z) - f(\zeta)| - |f(\zeta) - w| \geq \rho\delta/2 - \rho\delta/4 = \rho\delta/4.$$

Since $\phi(\zeta) < \rho\delta/4$, it follows that ϕ attains its minimum value in B at an interior point z_0. Since $z_0 \notin E$, we can take

$$z = z_0 - h[f'(z_0)]^{-1}\{f(z_0) - w\},$$

where $h > 0$ is so small that $|z - \zeta| < \delta$. Then

$$f(z) - w = f(z_0) - w + f'(z_0)(z - z_0) + o(h) = (1 - h)\{f(z_0) - w\} + o(h).$$

If $f(z_0) \neq w$ then, for sufficiently small $h > 0$,

$$|f(z) - w| \leq (1 - h/2)|f(z_0) - w| < |f(z_0) - w|,$$

which contradicts the definition of z_0. We conclude that $f(z_0) = w$. Thus $f(G \backslash E)$ contains not only $f(\zeta)$, but also an open disc $\{w \in \mathbb{C}: |w - f(\zeta)| < \rho\delta/4\}$ surrounding it. Since this holds for every $\zeta \in G \backslash E$, it follows that $f(G \backslash E)$ is an open set.

Now let $\zeta \in E$ and assume that $f(G)$ does not contain any open neighbourhood of $\omega := f(\zeta)$. Then $f(z) \neq \omega$ for every $z \in G \backslash E$. Choose $\delta > 0$ so small that the closed

disc $B = \{z \in \mathbb{C}: |z-\zeta| \leq \delta\}$ is contained in G and contains no point of E except ζ. If $S = \{z \in \mathbb{C}: |z-\zeta| = \delta\}$ is the boundary of B, there exists an open disc U with centre ω that contains no point of $f(S)$. It follows that if $A = \{z \in \mathbb{C}: 0 < |z-\zeta| < \delta\}$ is the annulus $B\backslash(S \cup \{\zeta\})$, then $U\backslash\{\omega\}$ is the union of the disjoint nonempty open sets $U \cap \{\mathbb{C}\backslash f(B)\}$ and $U \cap f(A)$. Since $U\backslash\{\omega\}$ is a connected set (because it is *path-connected*), this is a contradiction. □

From Proposition 29 we readily obtain

Theorem 30 *If*

$$f(z) = z^n + a_1 z^{n-1} + \cdots + a_n$$

is a polynomial of degree $n \geq 1$ *with complex coefficients* $a_1, \ldots, a_n$, *then* $f(\zeta) = 0$ *for some* $\zeta \in \mathbb{C}$.

Proof Since

$$f(z)/z^n = 1 + a_1/z + \cdots + a_n/z^n \to 1 \quad \text{as } |z| \to \infty,$$

we can choose $R > 0$ so large that

$$|f(z)| > |f(0)| \quad \text{for all } z \in \mathbb{C} \text{ such that } |z| = R.$$

Since the closed disc $D = \{z \in \mathbb{C}: |z| \leq R\}$ is compact, the continuous function $|f(z)|$ assumes its minimum value in D at a point ζ in the interior $G = \{z \in \mathbb{C}: |z| < R\}$. The function $f(z)$ is differentiable in G and the set E of all points of G at which the derivative $f'(z)$ vanishes is finite. (In fact E contains at most $n-1$ points, by Proposition II.15.) Hence, by Proposition 29, $f(G)$ is an open subset of $\mathbb{C}$. Since $|f(z)| \geq |f(\zeta)|$ for all $z \in G$, this implies $f(\zeta) = 0$. □

The first 'proof' of the fundamental theorem of algebra was given by d'Alembert (1746). Assuming the convergence of what are now called Puiseux expansions, he showed that if a polynomial assumes a value $w \neq 0$, then it also assumes a value w' such that $|w'| < |w|$. A much simpler way of reaching this conclusion, which required only the existence of k-th roots of complex numbers, was given by Argand (1814). Cauchy (1820) gave a similar proof and, with latter-day rigour, it is still reproduced in textbooks. The proof we have given rests on the same general principle, but uses neither the existence of k-th roots nor the continuity of the derivative. These may be called *differential calculus proofs*.

The basis for an *algebraic proof* was given by Euler (1749). His proof was completed by Lagrange (1772) and then simplified by Laplace (1795). The algebraic proof starts from the facts that $\mathbb{R}$ is an ordered field, that any positive element of $\mathbb{R}$ has a square root in $\mathbb{R}$ and that any polynomial of odd degree with coefficients from $\mathbb{R}$ has a root in $\mathbb{R}$. It then shows that any polynomial of degree $n \geq 1$ with coefficients from $\mathbb{C} = \mathbb{R}(i)$, where $i^2 = -1$, has a root in $\mathbb{C}$ by using induction on the highest power of 2 which divides n.

Gauss (1799) objected to this proof, because it assumed that there were 'roots' and then proved that these roots were complex numbers. The difficulty disappears if one

uses the result, due to Kronecker (1887), that a polynomial with coefficients from an arbitrary field K decomposes into linear factors in a field L which is a finite extension of K. This general result, which is not difficult to prove, is actually all that is required for many of the previous applications of the fundamental theorem of algebra.

It is often said that the first rigorous proof of the fundamental theorem of algebra was given by Gauss (1799). Like d'Alembert, however, Gauss assumed properties of algebraic curves which were unknown at the time. The gaps in this proof of Gauss were filled by Ostrowski (1920).

There are also *topological proofs* of the fundamental theorem of algebra, e.g. using the notion of topological degree. This type of proof is intuitively appealing, but not so easy to make rigorous. Finally, there are *complex analysis proofs*, which depend ultimately on Cauchy's theorem on complex line integrals. (The latter proofs are more closely related to either the differential calculus proofs or the topological proofs than they seem to be at first sight.)

The *exponential function* e^z may be defined, for any complex value of z, as the sum of the everywhere convergent power series

$$\sum_{n\geq 0} z^n/n! = 1 + z + z^2/2! + z^3/3! + \cdots .$$

It is easily verified that $w(z) = e^z$ is a solution of the differential equation $dw/dz = w$ satisfying the initial condition $w(0) = 1$.

For any $\zeta \in \mathbb{C}$, put $\varphi(z) = e^{\zeta - z}e^z$. Differentiating by the product rule, we obtain

$$\varphi'(z) = -e^{\zeta - z}e^z + e^{\zeta - z}e^z = 0.$$

Since this holds for all $z \in \mathbb{C}$, $\varphi(z)$ is a constant. Thus $\varphi(z) = \varphi(0) = e^\zeta$. Replacing ζ by $\zeta + z$, we obtain the *addition theorem* for the exponential function:

$$e^\zeta e^z = e^{\zeta + z} \quad \text{for all } z, \zeta \in \mathbb{C}.$$

In particular, $e^{-z}e^z = 1$ and hence $e^z \neq 0$ for every $z \in \mathbb{C}$.

The power series for e^z shows that, for any real y, e^{-iy} is the complex conjugate of e^{iy} and hence

$$|e^{iy}|^2 = e^{iy}e^{-iy} = 1.$$

It follows that, for all real x, y,

$$|e^{x+iy}| = |e^x||e^{iy}| = e^x.$$

The *trigonometric functions* $\cos z$ and $\sin z$ may be defined, for any complex value of z, by the formulas of Euler (1740):

$$\cos z = (e^{iz} + e^{-iz})/2, \quad \sin z = (e^{iz} - e^{-iz})/2i.$$

It follows at once that

$$\begin{aligned} e^{iz} &= \cos z + i \sin z, \\ \cos 0 = 1, \quad \sin 0 &= 0, \\ \cos(-z) = \cos z, \quad \sin(-z) &= -\sin z, \end{aligned}$$

and the relation $e^{iz}e^{-iz} = 1$ implies that

$$\cos^2 z + \sin^2 z = 1.$$

From the power series for e^z we obtain, for every $z \in \mathbb{C}$,

$$\cos z = \sum_{n\geq 0}(-1)^n z^{2n}/(2n)! = 1 - z^2/2! + z^4/4! - \cdots,$$
$$\sin z = \sum_{n\geq 0}(-1)^n z^{2n+1}/(2n+1)! = z - z^3/3! + z^5/5! - \cdots.$$

From the differential equation we obtain, for every $z \in \mathbb{C}$,

$$d(\cos z)/dz = -\sin z, \quad d(\sin z)/dz = \cos z.$$

From the addition theorem we obtain, for all $z, \zeta \in \mathbb{C}$,

$$\cos(z+\zeta) = \cos z \cos\zeta - \sin z \sin\zeta,$$
$$\sin(z+\zeta) = \sin z \cos\zeta + \cos z \sin\zeta.$$

We now consider periodicity properties. By the addition theorem for the exponential function, $e^{z+h} = e^z$ if and only if $e^h = 1$. Thus the exponential function has period h if and only if $e^h = 1$. Since $e^h = 1$ implies $h = ix$ for some real x, and since $\cos x$ and $\sin x$ are real for real x, the periods correspond to those real values of x for which

$$\cos x = 1, \quad \sin x = 0.$$

In fact, the second relation follows from the first, since $\cos^2 x + \sin^2 x = 1$.

By bracketing the power series for $\cos x$ in the form

$$\cos x = (1 - x^2/2! + x^4/4!) - (1 - x^2/7\cdot 8)x^6/6! - (1 - x^2/11\cdot 12)x^{10}/10! - \cdots$$

and taking $x = 2$, we see that $\cos 2 < 0$. Since $\cos 0 = 1$ and $\cos x$ is a continuous function of x, there is a least positive value ξ of x such that $\cos\xi = 0$. Then $\sin^2\xi = 1$. In fact $\sin\xi = 1$, since $\sin 0 = 0$ and $\sin' x = \cos x > 0$ for $0 \leq x < \xi$. Thus

$$0 < \sin x < 1 \quad \text{for } 0 < x < \xi$$

and

$$e^{i\xi} = \cos\xi + i\sin\xi = i.$$

As usual, we now write $\pi = 2\xi$. From $e^{\pi i/2} = i$, we obtain

$$e^{2\pi i} = i^4 = (-1)^2 = 1.$$

Thus the exponential function has period $2\pi i$. It follows that it also has period $2n\pi i$, for every $n \in \mathbb{Z}$. We will show that there are no other periods.

Suppose $e^{ix'} = 1$ for some $x' \in \mathbb{R}$ and choose $n \in \mathbb{Z}$ so that $n \leq x'/2\pi < n + 1$. If $x = x' - 2n\pi$, then $e^{ix} = 1$ and $0 \leq x < 2\pi$. If $x \neq 0$, then $0 < x/4 < \pi/2$ and hence $0 < \sin x/4 < 1$. Thus $e^{ix/4} \neq \pm 1, \pm i$. But this is a contradiction, since

$$(e^{ix/4})^4 = e^{ix} = 1.$$

We show next that the map $x \to e^{ix}$ maps the interval $0 \leq x < 2\pi$ bijectively onto the *unit circle*, i.e. the set of all complex numbers w such that $|w| = 1$. We already know that $|e^{ix}| = 1$ if $x \in \mathbb{R}$. If $e^{ix} = e^{ix'}$, where $0 \leq x \leq x' < 2\pi$, then $e^{i(x'-x)} = 1$. Since $0 \leq x' - x < 2\pi$, this implies $x' = x$.

It remains to show that if $u, v \in \mathbb{R}$ and $u^2 + v^2 = 1$, then

$$u = \cos x, \quad v = \sin x$$

for some x such that $0 \leq x < 2\pi$. If $u, v > 0$, then also $u, v < 1$. Hence $u = \cos x$ for some x such that $0 < x < \pi/2$. It follows that $v = \sin x$, since $\sin^2 x = 1 - u^2 = v^2$ and $\sin x > 0$. The other possible sign combinations for u, v may be reduced to the case $u, v > 0$ by means of the relations

$$\sin(x + \pi/2) = \cos x, \quad \cos(x + \pi/2) = -\sin x.$$

If z is any nonzero complex number, then $r = |z| > 0$ and $|z/r| = 1$. It follows that any nonzero complex number z can be uniquely expressed in the form

$$z = re^{i\theta},$$

where r, θ are real numbers such that $r > 0$ and $0 \leq \theta < 2\pi$. We call these r, θ the *polar coordinates* of z and θ the *argument* of z. If $z = x + iy$, where $x, y \in \mathbb{R}$, then $r = \sqrt{(x^2 + y^2)}$ and

$$x = r\cos\theta, \quad y = r\sin\theta.$$

Hence, in the geometrical representation of complex numbers by points of a plane, r is the distance of z from O and θ measures the angle between the positive x-axis and the ray $\overrightarrow{Oz}$.

We now show that the exponential function assumes every nonzero complex value w. Since $|w| > 0$, we have $|w| = e^x$ for some $x \in \mathbb{R}$. If $w' = w/|w|$, then $|w'| = 1$ and so $w' = e^{iy}$ for some $y \in \mathbb{R}$. Consequently,

$$w = |w|w' = e^x e^{iy} = e^{x+iy}.$$

It follows that, for any positive integer n, a nonzero complex number w has n distinct n-th roots. In fact, if $w = e^z$, then w has the distinct n-th roots

$$\zeta_k = \zeta\omega^k (k = 0, 1, \ldots, n-1),$$

where $\zeta = e^{z/n}$ and $\omega = e^{2\pi i/n}$. In the geometrical representation of complex numbers by points of a plane, the n-th roots of w are the vertices of an n-sided regular polygon.

It remains to show that π has its usual geometric significance. Since the continuously differentiable function $z(t) = e^{it}$ describes the unit circle as t increases from 0 to 2π, the length of the unit circle is

$$L = \int_0^{2\pi} |z'(t)| dt.$$

But $|z'(t)| = 1$, since $z'(t) = ie^{it}$, and hence $L = 2\pi$.

In a course of complex analysis one would now define complex line integrals, prove Cauchy's theorem and deduce its numerous consequences. The miracle is that, if $D = \{z \in \mathbb{C}: |z| < \rho\}$ is a disc with centre the origin, then any differentiable function $f: D \to \mathbb{C}$ can be represented by a *power series*,

$$f(z) = c_0 + c_1 z + c_2 z^2 + \cdots,$$

which is convergent for $|z| < \rho$. It follows that, if f vanishes at a sequence of distinct points converging to 0, then it vanishes everywhere. This is the basis for *analytic continuation*.

A complex-valued function f is said to be *holomorphic* at $a \in \mathbb{C}$ if, in some neighbourhood of a, it can be represented as the sum of a convergent power series (its 'Taylor' series):

$$f(z) = c_0 + c_1(z-a) + c_2(z-a)^2 + \cdots.$$

It is said to be *meromorphic* at $a \in \mathbb{C}$ if, for some integer n, it can be represented near a as the sum of a convergent series (its 'Laurent' series):

$$f(z) = c_0(z-a)^{-n} + c_1(z-a)^{-n+1} + c_2(z-a)^{-n+2} + \cdots.$$

If $c_0 \neq 0$, then $(z-a)f'(z)/f(z) \to -n$ as $z \to a$. If also $n > 0$ we say that a is a *pole* of f of *order* n with *residue* c_{n-1}. If $n = 1$, the residue is c_0 and the pole is said to be *simple*.

Let G be a nonempty connected open subset of $\mathbb{C}$. From what has been said, if $f: G \to \mathbb{C}$ is differentiable throughout G, then it is also holomorphic throughout G. If f_1 and f_2 are holomorphic throughout G and f_2 is not identically zero, then the quotient $f = f_1/f_2$ is meromorphic throughout G. Conversely, it may be shown that if f is meromorphic throughout G, then $f = f_1/f_2$ for some functions f_1, f_2 which are holomorphic throughout G.

The behaviour of many functions is best understood by studying them in the complex domain, as the exponential and trigonometric functions already illustrate. Complex numbers, when they first appeared, were called 'impossible' numbers. They are now indispensable.

6 Quaternions and Octonions

Quaternions were invented by Hamilton (1843) in order to be able to 'multiply' points of 3-dimensional space, in the same way that complex numbers enable one to multiply

points of a plane. The definition of quaternions adopted here will be analogous to our definition of complex numbers.

We define a *quaternion* to be a 2×2 matrix of the form

$$A = \begin{pmatrix} a & b \\ -\bar{b} & \bar{a} \end{pmatrix},$$

where a and b are complex numbers and the bar denotes complex conjugation. The set of all quaternions will be denoted by $\mathbb{H}$. We may define addition and multiplication in $\mathbb{H}$ to be matrix addition and multiplication, since $\mathbb{H}$ is closed under these operations. Furthermore $\mathbb{H}$ contains

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

and $A \in \mathbb{H}$ implies $-A \in \mathbb{H}$.

It follows from the properties of matrix addition and multiplication that addition and multiplication of quaternions have the properties **(A2)–(A5)** and **(M3)–(M4)**, with 0 and 1 as identity elements for addition and multiplication respectively. However, **(M2)** no longer holds, since multiplication is not always commutative. For example,

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \neq \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

On the other hand, there are now two distributive laws: *for all* $A, B, C \in \mathbb{H}$,

$$A(B + C) = AB + AC, \quad (B + C)A = BA + CA.$$

It is easily seen that $A \in \mathbb{H}$ is in the *centre* of $\mathbb{H}$, i.e. $AB = BA$ for every $B \in \mathbb{H}$, if and only if $A = \lambda 1$ for some real number λ. Since the map $\lambda \to \lambda 1$ preserves sums and products, we can regard $\mathbb{R}$ as contained in $\mathbb{H}$ by identifying the real number λ with the quaternion $\lambda 1$.

We define the *conjugate* of the quaternion

$$A = \begin{pmatrix} a & b \\ -\bar{b} & \bar{a} \end{pmatrix},$$

to be the quaternion

$$\bar{A} = \begin{pmatrix} \bar{a} & -b \\ \bar{b} & a \end{pmatrix}.$$

It is easily verified that

$$\overline{A + B} = \bar{A} + \bar{B}, \quad \overline{AB} = \bar{B}\bar{A}, \quad \bar{\bar{A}} = A.$$

Furthermore,

$$\bar{A}A = A\bar{A} = n(A), \quad A + \bar{A} = t(A),$$

where the *norm* $n(A)$ and *trace* $t(A)$ are both real:

$$n(A) = a\bar{a} + b\bar{b}, \quad t(A) = a + \bar{a}.$$

Moreover, $n(A) > 0$ if $A \neq 0$. It follows that any quaternion $A \neq 0$ has a multiplicative inverse: if $A^{-1} = n(A)^{-1}\bar{A}$, then

$$A^{-1}A = AA^{-1} = 1.$$

Norms and traces have the following properties: *for all* $A, B \in \mathbb{H}$,

$$\begin{aligned} t(\bar{A}) &= t(A), \\ n(\bar{A}) &= n(A), \\ t(A+B) &= t(A) + t(B), \\ n(AB) &= n(A)n(B). \end{aligned}$$

Only the last property is not immediately obvious, and it can be proved in one line:

$$n(AB) = \overline{AB}AB = \bar{B}\bar{A}AB = n(A)\bar{B}B = n(A)n(B).$$

Furthermore, for any $A \in \mathbb{H}$ we have

$$A^2 - t(A)A + n(A) = 0,$$

since the left side can be written in the form $A^2-(A+\bar{A})A+\bar{A}A$. (The relation is actually just a special case of the 'Cayley–Hamilton theorem' of linear algebra.) It follows that the quadratic polynomial x^2+1 has not two, but infinitely many quaternionic roots.

If we put

$$I = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad J = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}, \quad K = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix},$$

then

$$\begin{aligned} I^2 = J^2 = K^2 = -1, \\ IJ = K = -JI, \quad JK = I = -KJ, \quad KI = J = -IK. \end{aligned}$$

Moreover, any quaternion A can be uniquely represented in the form

$$A = \alpha_0 + \alpha_1 I + \alpha_2 J + \alpha_3 K,$$

where $\alpha_0, \ldots, \alpha_3 \in \mathbb{R}$. In fact this is equivalent to the previous representation with

$$a = \alpha_0 + i\alpha_3, \quad b = \alpha_1 + i\alpha_2.$$

The corresponding representation of the conjugate quaternion is

$$\bar{A} = \alpha_0 - \alpha_1 I - \alpha_2 J - \alpha_3 K.$$

Hence $\bar{A} = A$ if and only if $\alpha_1 = \alpha_2 = \alpha_3 = 0$ and $\bar{A} = -A$ if and only if $\alpha_0 = 0$.

A quaternion A is said to be *pure* if $\bar{A} = -A$. Thus any quaternion can be uniquely represented as the sum of a real number and a pure quaternion.

It follows from the multiplication table for the units I, J, K that $A = \alpha_0 + \alpha_1 I + \alpha_2 J + \alpha_3 K$ has norm

$$n(A) = \alpha_0^2 + \alpha_1^2 + \alpha_2^2 + \alpha_3^2.$$

Consequently the relation $n(A)n(B) = n(AB)$ may be written in the form

$$(\alpha_0^2 + \alpha_1^2 + \alpha_2^2 + \alpha_3^2)(\beta_0^2 + \beta_1^2 + \beta_2^2 + \beta_3^2) = \gamma_0^2 + \gamma_1^2 + \gamma_2^2 + \gamma_3^2,$$

where

$$\begin{aligned}
\gamma_0 &= \alpha_0\beta_0 - \alpha_1\beta_1 - \alpha_2\beta_2 - \alpha_3\beta_3,\\
\gamma_1 &= \alpha_0\beta_1 + \alpha_1\beta_0 + \alpha_2\beta_3 - \alpha_3\beta_2,\\
\gamma_2 &= \alpha_0\beta_2 - \alpha_1\beta_3 + \alpha_2\beta_0 + \alpha_3\beta_1,\\
\gamma_3 &= \alpha_0\beta_3 + \alpha_1\beta_2 - \alpha_2\beta_1 + \alpha_3\beta_0.
\end{aligned}$$

This '4-squares identity' was already known to Euler (1770).

An important application of quaternions is to the parametrization of rotations in 3-dimensional space. In describing this application it will be convenient to denote quaternions now by lower case letters. In particular, we will write i, j, k in place of I, J, K.

Let u be a quaternion with norm $n(u) = 1$, and consider the mapping $T : \mathbb{H} \to \mathbb{H}$ defined by

$$Tx = uxu^{-1}.$$

Evidently

$$\begin{aligned}
T(x + y) &= Tx + Ty,\\
T(xy) &= (Tx)(Ty),\\
T(\lambda x) &= \lambda Tx \quad \text{if } \lambda \in \mathbb{R}.
\end{aligned}$$

Moreover, since $u^{-1} = \bar{u}$,

$$T\bar{x} = \overline{Tx}.$$

It follows that

$$n(Tx) = n(x),$$

since

$$n(Tx) = Tx\overline{Tx} = TxT\bar{x} = T(x\bar{x}) = n(x)T1 = n(x).$$

Furthermore, T maps pure quaternions into pure quaternions, since $\bar{x} = -x$ implies

$$\overline{Tx} = T\bar{x} = -Tx.$$

If we write

$$x = \xi_1 i + \xi_2 j + \xi_3 k,$$

then

$$Tx = y = \eta_1 i + \eta_2 j + \eta_3 k,$$

where $\eta_\mu = \sum_{\upsilon=1}^{3} \beta_{\mu\upsilon}\xi_\upsilon$ for some $\beta_{\mu\upsilon} \in \mathbb{R}$. Since

$$\eta_1^2 + \eta_2^2 + \eta_3^2 = \xi_1^2 + \xi_2^2 + \xi_3^2,$$

the matrix $V = (\beta_{\mu\upsilon})$ is *orthogonal*: $V^{-1} = V^t$.

Thus with every quaternion u with norm 1 there is associated a 3×3 orthogonal matrix $V = (\beta_{\mu\upsilon})$. Explicitly, if

$$u = \alpha_0 + \alpha_1 i + \alpha_2 j + \alpha_3 k,$$

where

$$\alpha_0^2 + \alpha_1^2 + \alpha_2^2 + \alpha_3^2 = 1,$$

then

$$\begin{aligned}
&\beta_{11} = \alpha_0^2 + \alpha_1^2 - \alpha_2^2 - \alpha_3^2, \quad \beta_{12} = 2(\alpha_1\alpha_2 - \alpha_0\alpha_3), \quad \beta_{13} = 2(\alpha_1\alpha_3 + \alpha_0\alpha_2),\\
&\beta_{21} = 2(\alpha_1\alpha_2 + \alpha_0\alpha_3), \quad \beta_{22} = \alpha_0^2 - \alpha_1^2 + \alpha_2^2 - \alpha_3^2, \quad \beta_{23} = 2(\alpha_2\alpha_3 - \alpha_0\alpha_1),\\
&\beta_{31} = 2(\alpha_1\alpha_3 - \alpha_0\alpha_2), \quad \beta_{32} = 2(\alpha_2\alpha_3 + \alpha_0\alpha_1), \quad \beta_{33} = \alpha_0^2 - \alpha_1^2 - \alpha_2^2 + \alpha_3^2.
\end{aligned}$$

This parametrization of orthogonal transformations was first discovered by Euler(1770).

We now consider the dependence of V on u, and consequently write $V(u)$ in place of V. Since

$$u_1 u_2 x (u_1 u_2)^{-1} = u_1 (u_2 x u_2^{-1}) u_1^{-1},$$

we have

$$V(u_1 u_2) = V(u_1) V(u_2).$$

Thus the map $u \to V(u)$ is a 'homomorphism' of the multiplicative group of all quaternions of norm 1 into the group of all 3×3 real orthogonal matrices. In particular, $V(\bar{u}) = V(u)^{-1}$.

We show next that two quaternions u_1, u_2 of norm 1 yield the same orthogonal matrix if and only if $u_2 = \pm u_1$. Put $u = u_2^{-1} u_1$. Then $u_1 x u_1^{-1} = u_2 x u_2^{-1}$ if and only if $ux = xu$. This holds for every pure quaternion x if and only if u is real, i.e. if and only if $u = \pm 1$, since $n(u) = 1$.

The question arises whether all 3×3 orthogonal matrices may be represented in the above way. It follows readily from the preceding formulas for $\beta_{\mu\upsilon}$ that the orthogonal matrix $-I$ cannot be so represented. Consequently, if an orthogonal matrix V is

represented, then $-V$ is not. On the other hand, suppose u is a pure quaternion, so that $\alpha_0 = 0$. Then $ux + xu = ux + \bar{x}\bar{u}$ is real, and given by

$$ux + xu = -2(\alpha_1\xi_1 + \alpha_2\xi_2 + \alpha_3\xi_3) = 2\langle \bar{u}, x\rangle,$$

with the notation of §10 for inner products in $\mathbb{R}^3$. It follows that

$$y = ux\bar{u} = 2\langle \bar{u}, x\rangle\bar{u} - x.$$

But the mapping $x \to x - 2\langle \bar{u}, x\rangle\bar{u}$ is a *reflection* in the plane orthogonal to the unit vector u. Hence, for every reflection R, $-R$ is represented. It may be shown that every orthogonal transformation of $\mathbb{R}^3$ is a product of reflections. (Indeed, this is a special case of a more general result which will be proved in Proposition 17 of Chapter VII.) It follows that an orthogonal matrix V is represented if and only if V is a product of an even number of reflections (or, equivalently, if and only if V has determinant 1, as defined in Chapter V, §1).

Since, by our initial definition of quaternions, the quaternions of norm 1 are just the 2×2 unitary matrices with determinant 1, our results may be summed up (cf. Chapter X, §8) by saying that there is a homomorphism of the *special unitary group* $SU_2(\mathbb{C})$ onto the *special orthogonal group* $SO_3(\mathbb{R})$, with kernel $\{\pm I\}$. (Here 'special' signifies 'determinant 1'.)

Since the quaternions of norm 1 may be identified with the points of the unit sphere S^3 in $\mathbb{R}^4$ it follows that, as a topological space, $SO_3(\mathbb{R})$ is homeomorphic to S^3 with antipodal points identified, i.e. to the projective space $P^3(\mathbb{R})$. Similarly (cf. Chapter X, §8), the topological group $SU_2(\mathbb{C})$ is the *simply-connected covering space* of the topological group $SO_3(\mathbb{R})$.

Again, by considering the map $T: \mathbb{H} \to \mathbb{H}$ defined by $Tx = vxu^{-1}$, where u, v are quaternions with norm 1, it may be seen that that there is a homomorphism of the direct product $SU_2(\mathbb{C}) \times SU_2(\mathbb{C})$ onto the special orthogonal group $SO_4(\mathbb{R})$ of 4×4 real orthogonal matrices with determinant 1, the kernel being $\{\pm(I, I)\}$.

Almost immediately after Hamilton's invention of quaternions Graves (1844), in a letter to Hamilton, and Cayley (1845) invented 'octonions', also known as 'octaves' or 'Cayley numbers'. We define an *octonion* to be an ordered pair (a_1, a_2) of quaternions, with addition and multiplication defined by

$$(a_1, a_2) + (b_1, b_2) = (a_1 + b_1, a_2 + b_2),$$
$$(a_1, a_2) \cdot (b_1, b_2) = (a_1 b_1 - \bar{b_2} a_2, b_2 a_1 + a_2 \bar{b_1}).$$

Then the set $\mathbb{O}$ of all octonions is a commutative group under addition, i.e. the laws **(A2)–(A5)** hold, with $0 = (0, 0)$ as identity element, and multiplication is both left and right distributive with respect to addition. The octonion $1 = (1, 0)$ is a two-sided identity element for multiplication, and the octonion $\varepsilon = (0, 1)$ has the property $\varepsilon^2 = -1$.

It is easily seen that $\alpha \in \mathbb{O}$ is in the *centre* of $\mathbb{O}$, i.e. $\alpha\beta = \beta\alpha$ for every $\beta \in \mathbb{O}$, if and only if $\alpha = (c, 0)$ for some $c \in \mathbb{R}$.

Since the map $a \to (a, 0)$ preserves sums and products, we may regard $\mathbb{H}$ as contained in $\mathbb{O}$ by identifying the quaternion a with the octonion $(a, 0)$. This shows that multiplication of octonions is in general not commutative. It is also in general not even associative; for example,

$$(ij)\varepsilon = k\varepsilon = (0, k), \quad i(j\varepsilon) = i(0, j) = (0, -k).$$

It is for this reason that we defined octonions as ordered pairs, rather than as matrices. It should be mentioned, however, that we could have used precisely the same construction to define complex numbers as ordered pairs of real numbers, and quaternions as ordered pairs of complex numbers, but the verification of the associative law for multiplication would then have been more laborious.

Although multiplication is non-associative, $\mathbb{O}$ does inherit some other properties from $\mathbb{H}$. If we define the *conjugate* of the octonion $\alpha = (a_1, a_2)$ to be the octonion $\bar{\alpha} = (\overline{a_1}, -a_2)$, then it is easily verified that

$$\overline{\alpha + \beta} = \bar{\alpha} + \bar{\beta}, \quad \overline{\alpha\beta} = \bar{\beta}\bar{\alpha}, \quad \bar{\bar{\alpha}} = \alpha.$$

Furthermore,

$$\alpha\bar{\alpha} = \bar{\alpha}\alpha = n(\alpha),$$

where the *norm* $n(\alpha) = a_1\overline{a_1} + a_2\overline{a_2}$ is real. Moreover $n(\alpha) > 0$ if $\alpha \neq 0$, and $n(\bar{\alpha}) = n(\alpha)$.

It will now be shown that if $\alpha, \beta \in \mathbb{O}$ and $\alpha \neq 0$, then the equation

$$\xi\alpha = \beta$$

has a unique solution $\xi \in \mathbb{O}$. Writing $\alpha = (a_1, a_2)$, $\beta = (b_1, b_2)$ and $\xi = (x_1, x_2)$, we have to solve the simultaneous quaternionic equations

$$\begin{aligned} x_1a_1 - \overline{a_2}x_2 &= b_1, \\ a_2x_1 + x_2\overline{a_1} &= b_2. \end{aligned}$$

If we multiply the second equation on the right by a_1 and replace x_1a_1 by its value from the first equation, we get

$$n(\alpha)x_2 = b_2a_1 - a_2b_1.$$

Similarly, if we multiply the first equation on the right by $\overline{a_1}$ and replace $x_2\overline{a_1}$ by its value from the second equation, we get

$$n(\alpha)x_1 = b_1\overline{a_1} + \overline{a_2}b_2.$$

It follows that the equation $\xi\alpha = \beta$ has the unique solution

$$\xi = n(\alpha)^{-1}\beta\bar{\alpha}.$$

Since the equation $\alpha\eta = \beta$ is equivalent to $\bar{\eta}\bar{\alpha} = \bar{\beta}$, it has the unique solution $\eta = n(\alpha)^{-1}\bar{\alpha}\beta$. Thus $\mathbb{O}$ is a *division algebra*. It should be noted that, since $\mathbb{O}$ is non-associative, it is not enough to verify that every nonzero element has a multiplicative inverse.

It follows from the preceding discussion that, *for all* $\alpha, \beta \in \mathbb{O}$,

$$(\beta\bar{\alpha})\alpha = n(\alpha)\beta = \alpha(\bar{\alpha}\beta).$$

Consequently the norm is multiplicative: *for all* $\alpha, \beta \in \mathbb{O}$,

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

For, putting $\gamma = \alpha\beta$, we have

$$n(\gamma)\bar{\alpha} = (\bar{\alpha}\gamma)\bar{\gamma} = (\bar{\alpha}(\alpha\beta))\bar{\gamma} = n(\alpha)\beta\bar{\gamma} = n(\alpha)\beta(\bar{\beta}\bar{\alpha}) = n(\alpha)n(\beta)\bar{\alpha}.$$

This establishes the result when $\alpha \neq 0$, and when $\alpha = 0$ it is obvious.

Every $\alpha \in \mathbb{O}$ has a unique representation $\alpha = a_1 + a_2\varepsilon$, where $a_1, a_2 \in \mathbb{H}$, and hence a unique representation

$$\alpha = c_0 + c_1 i + c_2 j + c_3 k + c_4\varepsilon + c_5 i\varepsilon + c_6 j\varepsilon + c_7 k\varepsilon,$$

where $c_0, \ldots, c_7 \in \mathbb{R}$. Since $\bar{\alpha} = \overline{a_1} - a_2\varepsilon$ and $n(\alpha) = a_1\overline{a_1} + a_2\overline{a_2}$, it follows that

$$\bar{\alpha} = c_0 - c_1 i - c_2 j - c_3 k - c_4\varepsilon - c_5 i\varepsilon - c_6 j\varepsilon - c_7 k\varepsilon$$

and

$$n(\alpha) = c_0^2 + \cdots + c_7^2.$$

Consequently the relation $n(\alpha)n(\beta) = n(\alpha\beta)$ may be written in the form

$$(c_0^2 + \cdots + c_7^2)(d_0^2 + \cdots + d_7^2) = e_0^2 + \cdots + e_7^2,$$

where $e_i = \sum_{j=0}^{7}\sum_{k=0}^{7} \rho_{ijk} c_j d_k$ for some real constants ρ_{ijk} which do not depend on the c's and d's. An '8-squares identity' of this type was first found by Degen (1818).

7 Groups

A nonempty set G is said to be a *group* if a binary operation φ, i.e. a mapping $\varphi: G \times G \to G$, is defined with the properties

(i) $\varphi(\varphi(a, b), c) = \varphi(a, \varphi(b, c))$ for all $a, b, c \in G$; (associative law)
(ii) there exists $e \in G$ such that $\varphi(e, a) = a$ for every $a \in G$; (identity element)
(iii) for each $a \in G$, there exists $a^{-1} \in G$ such that $\varphi(a^{-1}, a) = e$.(inverse elements)

If, in addition,

(iv) $\varphi(a, b) = \varphi(b, a)$ for all $a, b \in G$,(commutative law)

then the group G is said to be *commutative* or *abelian*.

For example, the set $\mathbb{Z}$ of all integers is a commutative group under addition, i.e. with $\varphi(a, b) = a + b$, with 0 as identity element and $-a$ as the inverse of a. Similarly, the set $\mathbb{Q}^\times$ of all nonzero rational numbers is a commutative group under multiplication, i.e. with $\varphi(a, b) = ab$, with 1 as identity element and a^{-1} as the inverse of a.

We now give an example of a noncommutative group. The set $\mathscr{S}_A$ of all bijective maps $f: A \to A$ of a nonempty set A to itself is a group under composition, i.e. with $\varphi(a, b) = a \circ b$, with the identity map i_A as identity element and the inverse map f^{-1} as the inverse of f. If A contains at least 3 elements, then $\mathscr{S}_A$ is a noncommutative

group. For suppose a, b, c are distinct elements of A, let $f: A \to A$ be the bijective map defined by

$$f(a) = b, \quad f(b) = a, \quad f(x) = x \quad \text{if } x \neq a, b,$$

and let $g: A \to A$ be the bijective map defined by

$$g(a) = c, \quad g(c) = a, \quad g(x) = x \quad \text{if } x \neq a, c.$$

Then $f \circ g \neq g \circ f$, since $(f \circ g)(a) = c$ and $(g \circ f)(a) = b$.

For arbitrary groups, instead of $\varphi(a, b)$ we usually write $a \cdot b$ or simply ab. For commutative groups, instead of $\varphi(a, b)$ we often write $a + b$.

Since, by the associative law,

$$(ab)c = a(bc),$$

we will usually dispense with brackets.

We now derive some simple properties possessed by all groups. By (iii) we have $a^{-1}a = e$. In fact also $aa^{-1} = e$. This may be seen by multiplying on the left, by the inverse of a^{-1}, the relation

$$a^{-1}aa^{-1} = ea^{-1} = a^{-1}.$$

By (ii) we have $ea = a$. It now follows that also $ae = a$, since

$$ae = aa^{-1}a = ea.$$

For all elements a, b of the group G, the equation $ax = b$ has the solution $x = a^{-1}b$ and the equation $ya = b$ has the solution $y = ba^{-1}$. Moreover, these solutions are unique. For from $ax = ax'$ we obtain $x = x'$ by multiplying on the left by a^{-1}, and from $ya = y'a$ we obtain $y = y'$ by multiplying on the right by a^{-1}.

In particular, the identity element e is unique, since it is the solution of $ea = a$, and the inverse a^{-1} of a is unique, since it is the solution of $a^{-1}a = e$. It follows that the inverse of a^{-1} is a and the inverse of ab is $b^{-1}a^{-1}$.

As the preceding argument suggests, in the definition of a group we could have replaced left identity and left inverse by right identity and right inverse, i.e. we could have required $ae = a$ and $aa^{-1} = e$, instead of $ea = a$ and $a^{-1}a = e$. (However, left identity and right inverse, or right identity and left inverse, would not give the same result.)

If H, K are nonempty subsets of a group G, we denote by HK the subset of G consisting of all elements hk, where $h \in H$ and $k \in K$. If L is also a nonempty subset of G, then evidently

$$(HK)L = H(KL).$$

A subset H of a group G is said to be a *subgroup* of G if it is a group under the same group operation as G itself. A nonempty subset H is a subgroup of G if and only if $a, b \in H$ implies $ab^{-1} \in H$. Indeed the necessity of the condition is obvious. It is also sufficient, since it implies first $e = aa^{-1} \in H$ and then $b^{-1} = eb^{-1} \in H$. (The associative law in H is inherited from G.)

We now show that a nonempty *finite* subset H of a group G is a subgroup of G if it is closed under multiplication only. For, if $a \in H$, then $ha \in H$ for all $h \in H$. Since H is finite and the mapping $h \to ha$ of H into itself is injective, it is also surjective by the pigeonhole principle (Corollary I.6). Hence $ha = a$ for some $h \in H$, which shows that H contains the identity element of G. It now further follows that $ha = e$ for some $h \in H$, which shows that H is also closed under inversion.

A group is said to be *finite* if it contains only finitely many elements and to be of *order n* if it contains exactly n elements.

In order to give an important example of a subgroup we digress briefly. Let n be a positive integer and let A be the set $\{1, 2, \ldots, n\}$ with the elements in their natural order. Since we regard A as ordered, a bijective map $\alpha: A \to A$ will be called a *permutation*. The set of all permutations of A is a group under composition, the *symmetric group* $\mathscr{S}_n$.

Suppose now that $n > 1$. An inversion of order induced by the permutation α is a pair (i, j) with $i < j$ for which $\alpha(i) > \alpha(j)$. The permutation α is said to be *even* or *odd* according as the total number of inversions of order is even or odd. For example, the permutation $\{1, 2, 3, 4, 5\} \to \{3, 5, 4, 1, 2\}$ is odd, since there are $2 + 3 + 2 = 7$ inversions of order.

The *sign* of the permutation α is defined by

$$\operatorname{sgn}(\alpha) = 1 \text{ or } -1 \text{ according as } \alpha \text{ is even or odd.}$$

Evidently we can write

$$\operatorname{sgn}(\alpha) = \prod_{1 \le i < j \le n} \{\alpha(j) - \alpha(i)\}/(j - i),$$

from which it follows that

$$\operatorname{sgn}(\alpha\beta) = \operatorname{sgn}(\alpha)\operatorname{sgn}(\beta).$$

Since the sign of the identity permutation is 1, this implies

$$\operatorname{sgn}(\alpha^{-1}) = \operatorname{sgn}(\alpha).$$

Thus $\operatorname{sgn}(\rho^{-1}\alpha\rho) = \operatorname{sgn}(\alpha)$ for any permutation ρ of A, and so $\operatorname{sgn}(\alpha)$ is actually independent of the ordering of A.

Since the product of two even permutations is again an even permutation, the even permutations form a subgroup of $\mathscr{S}_n$, the *alternating group* $\mathscr{A}_n$. The order of $\mathscr{A}_n$ is $n!/2$. For let τ be the permutation $\{1, 2, 3, \ldots, n\} \to \{2, 1, 3, \ldots, n\}$. Since there is only one inversion of order, τ is odd. Since $\tau\tau$ is the identity permutation, a permutation is odd if and only if it has the form $\alpha\tau$, where α is even. Hence the number of odd permutations is equal to the number of even permutations.

It may be mentioned that the sign of a permutation can also be determined without actually counting the total number of inversions. In fact any $\alpha \in \mathscr{S}_n$ may be written as a product of v disjoint cycles, and α is even or odd according as $n - v$ is even or odd.

We now return to the main story. Let H be a subgroup of an arbitrary group G and let a, b be elements of G. We write $a \sim_r b$ if $ba^{-1} \in H$. We will show that this is an equivalence relation.

The relation is certainly reflexive, since $e \in H$. It is also symmetric, since if $c = ba^{-1} \in H$, then $c^{-1} = ab^{-1} \in H$. Furthermore it is transitive, since if $ba^{-1} \in H$ and $cb^{-1} \in H$, then also $ca^{-1} = (cb^{-1})(ba^{-1}) \in H$.

The equivalence class which contains a is the set Ha of all elements ha, where $h \in H$. We call any such equivalence class a *right coset* of the subgroup H, and any element of a given coset is said to be a *representative* of that coset.

It follows from the remarks in §0 about arbitrary equivalence relations that, for any two cosets Ha and Ha', either $Ha = Ha'$ or $Ha \cap Ha' = \emptyset$. Moreover, the distinct right cosets form a partition of G.

If H is a subgroup of a finite group G, then H is also finite and the number of distinct right cosets is finite. Moreover, each right coset Ha contains the same number of elements as H, since the mapping $h \to ha$ of H to Ha is bijective. It follows that the order of the subgroup H divides the order of the whole group G, a result usually known as *Lagrange's theorem*. The quotient of the orders, i.e. the number of distinct cosets, is called the *index* of H in G.

Suppose again that H is a subgroup of an arbitrary group G and that $a, b \in G$. By writing $a \sim_l b$ if $a^{-1}b \in H$, we obtain another equivalence relation. The equivalence class which contains a is now the set aH of all elements ah, where $h \in H$. We call any such equivalence class a *left coset* of the subgroup H. Again, two left cosets either coincide or are disjoint, and the distinct left cosets form a partition of G.

When are the two partitions, into left cosets and into right cosets, the same? Evidently $Ha = aH$ for every $a \in G$ if and only if $a^{-1}Ha = H$ for every $a \in G$ or, since a may be replaced by a^{-1}, if and only if $a^{-1}ha \in H$ for every $h \in H$ and every $a \in G$. A subgroup H which satisfies this condition is said to be 'invariant' or *normal*.

Any group G obviously has two normal subgroups, namely G itself and the subset $\{e\}$ which contains only the identity element. A group G is said to be *simple* if it has no other normal subgroups and if these two are distinct (i.e., G contains more than one element).

We now show that if H is a normal subgroup of a group G, then the collection of all cosets of H can be given the structure of a group. Since $Ha = aH$ and $HH = H$, we have

$$(Ha)(Hb) = H(Ha)b = Hab.$$

Thus if we define the product $Ha \cdot Hb$ of the cosets Ha and Hb to be the coset Hab, the definition does not depend on the choice of coset representatives. Clearly multiplication of cosets is associative, the coset $H = He$ is an identity element and the coset Ha^{-1} is an inverse of the coset Ha. The new group thus constructed is called the *factor group* or *quotient group* of G by the normal subgroup H, and is denoted by G/H.

A mapping $f: G \to G'$ of a group G into a group G' is said to be a (group) *homomorphism* if

$$f(ab) = f(a)f(b) \quad \text{for all } a, b \in G.$$

By taking $a = b = e$, we see that this implies that $f(e) = e'$ is the identity element of G'. By taking $b = a^{-1}$, it now follows that $f(a^{-1})$ is the inverse of $f(a)$ in G'. Since the subset $f(G)$ of G' is closed under both multiplication and inversion, it is a subgroup of G'.

If $g: G' \to G''$ is a homomorphism of the group G' into a group G'', then the composite map $g \circ f: G \to G''$ is also a homomorphism.

The *kernel* of the homomorphism f is defined to be the set N of all $a \in G$ such that $f(a) = e'$ is the identity element of G'. The kernel is a subgroup of G, since if $a \in N$ and $b \in N$, then $ab \in N$ and $a^{-1} \in N$. Moreover, it is a normal subgroup, since $a \in N$ and $c \in G$ imply $c^{-1}ac \in N$.

For any $a \in G$, put $a' = f(a) \in G'$. The coset Na is the set of all $x \in G$ such that $f(x) = a'$, and the map $Na \to a'$ is a bijection from the collection of all cosets of N to $f(G)$. Since f is a homomorphism, Nab is mapped to $a'b'$. Hence the map $Na \to a'$ is a homomorphism of the factor group G/N to $f(G)$.

A mapping $f: G \to G'$ of a group G into a group G' is said to be a (group) *isomorphism* if it is both bijective and a homomorphism. The inverse mapping $f^{-1}: G' \to G$ is then also an isomorphism. (An *automorphism* of a group G is an isomorphism of G with itself.)

Thus we have shown that, if $f: G \to G'$ is a homomorphism of a group G into a group G', with kernel N, then the factor group G/N is isomorphic to $f(G)$.

Suppose now that G is an arbitrary group and a any element of G. We have already defined a^{-1}, the inverse of a. We now inductively define a^n, for any integer n, by putting

$$a^0 = e, \qquad a^1 = a,$$
$$a^n = a(a^{n-1}), \quad a^{-n} = a^{-1}(a^{-1})^{n-1} \quad \text{if } n > 1.$$

It is readily verified that, for all $m, n \in \mathbb{Z}$,

$$a^m a^n = a^{m+n}, \quad (a^m)^n = a^{mn}.$$

The set $\langle a \rangle = \{a^n : n \in \mathbb{Z}\}$ is a commutative subgroup of G, the *cyclic subgroup generated by* a. Evidently $\langle a \rangle$ contains a and is contained in every subgroup of G which contains a.

If we regard $\mathbb{Z}$ as a group under addition, then the mapping $n \to a^n$ is a homomorphism of $\mathbb{Z}$ onto $\langle a \rangle$. Consequently $\langle a \rangle$ is isomorphic to the factor group $\mathbb{Z}/N$, where N is the subgroup of $\mathbb{Z}$ consisting of all integers n such that $a^n = e$. Evidently $0 \in N$, and $n \in \mathbb{N}$ implies $-n \in N$. Thus either $N = \{0\}$ or N contains a positive integer. In the latter case, let s be the least positive integer in N. By Proposition 14, for any integer n there exist integers q, r such that

$$n = qs + r, \quad 0 \le r < s.$$

If $n \in N$, then also $r = n - qs \in N$ and hence $r = 0$, by the definition of s. It follows that $N = s\mathbb{Z}$ is the subgroup of $\mathbb{Z}$ consisting of all multiples of s. Thus either $\langle a \rangle$ is isomorphic to $\mathbb{Z}$, and is an infinite group, or $\langle a \rangle$ is isomorphic to the factor group $\mathbb{Z}/s\mathbb{Z}$, and is a finite group of order s. We say that the element a itself is of *infinite order* if $\langle a \rangle$ is infinite and of *order* s if $\langle a \rangle$ is of order s.

It is easily seen that in a *commutative* group the set of all elements of finite order is a subgroup, called its *torsion subgroup*.

If S is any nonempty subset of a group G, then the set $\langle S \rangle$ of all finite products $a_1^{\varepsilon_1} a_2^{\varepsilon_1} \cdots a_n^{\varepsilon_n}$, where $n \in \mathbb{N}$, $a_j \in S$ and $\varepsilon_j = \pm 1$, is a subgroup of G, called the

subgroup *generated* by S. Clearly $S \subseteq \langle S \rangle$ and $\langle S \rangle$ is contained in every subgroup of G which contains S.

Two elements a, b of a group G are said to be *conjugate* if $b = x^{-1}ax$ for some $x \in G$. It is easy to see that conjugacy is an equivalence relation. For $a = a^{-1}aa$, if $b = x^{-1}ax$ then $a = (x^{-1})^{-1}bx^{-1}$, and $b = x^{-1}ax, c = y^{-1}by$ together imply $c = (xy)^{-1}axy$. Consequently G may be partitioned into *conjugacy classes*, so that two elements of G are conjugate if and only if they belong to the same conjugacy class.

For any element a of a group G, the set N_a of all elements of G which commute with a,

$$N_a = \{x \in G : xa = ax\},$$

is closed under multiplication and inversion. Thus N_a is a subgroup of G, called the *centralizer* of a in G.

If y and z lie in the same right coset of N_a, so that $z = xy$ for some $x \in N_a$, then $zy^{-1}a = azy^{-1}$ and hence $y^{-1}ay = z^{-1}az$. Conversely, if $y^{-1}ay = z^{-1}az$, then y and z lie in the same right coset of N_a. If G is finite, it follows that the number of elements in the conjugacy class containing a is equal to the number of right cosets of the subgroup N_a, i.e. to the *index* of the subgroup N_a in G, and hence it divides the order of G.

To conclude, we mention a simple way of creating new groups from given ones. Let G, G' be groups and let $G \times G'$ be the set of all ordered pairs (a, a') with $a \in G$ and $a' \in G'$. Then $G \times G'$ acquires the structure of a group if we define the product $(a, a') \cdot (b, b')$ of (a, a') and (b, b') to be $(ab, a'b')$. Multiplication is clearly associative, (e, e') is an identity element and (a^{-1}, a'^{-1}) is an inverse for (a, a'). The group thus constructed is called the *direct product* of G and G', and is again denoted by $G \times G'$.

8 Rings and Fields

A nonempty set R is said to be a *ring* if two binary operations, $+$ (addition) and $\cdot$ (multiplication), are defined with the properties

(i) R is a commutative group under addition, with 0 (*zero*) as identity element and $-a$ as inverse of a;
(ii) multiplication is associative: $(ab)c = a(bc)$ for all $a, b, c \in R$;
(iii) there exists an identity element 1 (*one*) for multiplication: $a1 = a = 1a$ for every $a \in R$;
(iv) addition and multiplication are connected by the two distributive laws:

$$(a + b)c = (ac) + (bc), \quad c(a + b) = (ca) + (cb) \quad \text{for all } a, b, c \in R.$$

The elements 0 and 1 are necessarily uniquely determined. If, in addition, multiplication is commutative:

$$ab = ba \quad \text{for all } a, b \in R,$$

then R is said to be a *commutative* ring. In a commutative ring either one of the two distributive laws implies the other.

It may seem inconsistent to require that addition is commutative, but not multiplication. However, the commutative law for addition is actually a consequence of the other axioms for a ring. For, by the first distributive law we have

$$(a+b)(1+1) = a(1+1) + b(1+1) = a+a+b+b,$$

and by the second distributive law

$$(a+b)(1+1) = (a+b)1 + (a+b)1 = a+b+a+b.$$

Since a ring is a group under addition, by comparing these two relations we obtain first

$$a+a+b = a+b+a$$

and then $a+b = b+a$.

As examples, the set $\mathbb{Z}$ of all integers is a commutative ring, with the usual definitions of addition and multiplication, whereas if $n > 1$, the set $M_n(\mathbb{Z})$ of all $n \times n$ matrices with entries from $\mathbb{Z}$ is a noncommutative ring, with the usual definitions of matrix addition and multiplication.

A very different example is the collection $\mathscr{P}(X)$ of all subsets of a given set X. If we define the sum $A+B$ of two subsets A, B of X to be their *symmetric difference*, i.e. the set of all elements of X which are in either A or B, but not in both:

$$A+B = (A \cup B) \backslash (A \cap B) = (A \cup B) \cap (A^{\mathrm{c}} \cup B^{\mathrm{c}}),$$

and the product AB to be the set of all elements of X which are in both A and B:

$$AB = A \cap B,$$

it is not difficult to verify that $\mathscr{P}(X)$ is a commutative ring, with the empty set $\emptyset$ as identity element for addition and the whole set X as identity element for multiplication. For every $A \in \mathscr{P}(X)$, we also have

$$A+A = \emptyset, \quad AA = A.$$

The set operations are in turn determined by the ring operations:

$$A \cup B = A+B+AB, \quad A \cap B = AB, \quad A^{\mathrm{c}} = A+X.$$

A ring R is said to be a *Boolean ring* if $aa = a$ for every $a \in R$. It follows that $a+a = 0$ for every $a \in R$, since

$$a+a = (a+a)(a+a) = a+a+a+a.$$

Moreover, a Boolean ring is commutative, since

$$a+b = (a+b)(a+b) = a+b+ab+ba$$

and $ba = -ba$, by what we have already proved.

For an arbitrary set X, any nonempty subset of $\mathscr{P}(X)$ which is closed under union, intersection and complementation can be given the structure of a Boolean ring in the

manner just described. It was proved by Stone (1936) that every Boolean ring may be obtained in this way. Thus the algebraic laws of set theory may be replaced by the more familiar laws of algebra and all such laws are consequences of a small number among them.

We now return to arbitrary rings. In the same way as for $\mathbb{Z}$, in any ring R we have

$$a0 = 0 = 0a \quad \text{for every } a$$

and

$$(-a)b = -(ab) = a(-b) \quad \text{for all } a, b.$$

It follows that R contains only one element if $1 = 0$. We will say that the ring R is 'trivial' in this case.

Suppose R is a nontrivial ring. Then, viewing R as a group under addition, the cyclic subgroup $\langle 1 \rangle$ is either infinite, and isomorphic to $\mathbb{Z}/0\mathbb{Z}$, or finite of order s, and isomorphic to $\mathbb{Z}/s\mathbb{Z}$ for some positive integer s. The ring R is said to have *characteristic* 0 in the first case and *characteristic* s in the second case.

For any positive integer n, write

$$na := a + \cdots + a \quad (n \text{ summands}).$$

If R has characteristic $s > 0$, then $sa = 0$ for every $a \in R$, since

$$sa = (1 + \cdots + 1)a = 0a = 0.$$

On the other hand, $n1 \neq 0$ for every positive integer $n < s$, by the definition of characteristic.

An element a of a nontrivial ring R is said to be 'invertible' or a *unit* if there exists an element a^{-1} such that

$$a^{-1}a = 1 = aa^{-1}.$$

The element a^{-1} is then uniquely determined and is called the *inverse* of a. For example, 1 is a unit and is its own inverse. If a is a unit, then a^{-1} is also a unit and its inverse is a. If a and b are units, then ab is also a unit and its inverse is $b^{-1}a^{-1}$. It follows that the set $R^\times$ of all units is a group under multiplication.

A nontrivial ring R in which every nonzero element is invertible is said to be a *division ring*. Thus all nonzero elements of a division ring form a group under multiplication, the *multiplicative group* of the division ring. A *field* is a commutative division ring.

A nontrivial commutative ring R is said to be an *integral domain* if it has no 'divisors of zero', i.e. if $a \neq 0$ and $b \neq 0$ imply $ab \neq 0$. A division ring also has no divisors of zero, since if $a \neq 0$ and $b \neq 0$, then $a^{-1}ab = b \neq 0$, and hence $ab \neq 0$.

As examples, the set $\mathbb{Q}$ of rational numbers, the set $\mathbb{R}$ of real numbers and the set $\mathbb{C}$ of complex numbers are all fields, with the usual definitions of addition and multiplication. The set $\mathbb{H}$ of quaternions is a division ring, and the set $\mathbb{Z}$ of integers is an integral domain, but neither is a field.

In a ring with no divisors of zero, the additive order of any nonzero element a is the same as the additive order of 1, since $ma = (m1)a = 0$ if and only if $m1 = 0$. Furthermore, the characteristic of such a ring is either 0 or a prime number. For assume $n = lm$, where l and m are positive integers less than n. If $n1 = 0$, then

$$(l1)(m1) = n1 = 0.$$

Since there are no divisors of zero, either $l1 = 0$ or $m1 = 0$, and hence the characteristic cannot be n.

A subset S of a ring R is said to be a (two-sided) *ideal* if it is a subgroup of R under addition and if, for every $a \in S$ and $c \in R$, both $ac \in S$ and $ca \in S$.

Any ring R has two obvious ideals, namely R itself and the subset $\{0\}$. It is said to be *simple* if it has no other ideals and is nontrivial.

Any division ring is simple. For if an ideal S of a division ring R contains $a \neq 0$, then for every $c \in R$ we have $c = (ca^{-1})a \in S$.

Conversely, if a *commutative* ring R is simple, then it is a field. For, if a is any nonzero element of R, the set

$$S_a = \{xa : x \in R\}$$

is an ideal (since R is commutative). Since S_a contains $1a = a \neq 0$, we must have $S_a = R$. Hence $1 = xa$ for some $x \in R$. Thus every nonzero element of R is invertible.

If R is a commutative ring and $a_1, \ldots, a_m \in R$, then the set S consisting of all elements $x_1a_1 + \cdots + x_ma_m$, where $x_j \in R$ $(1 \leq j \leq m)$, is clearly an ideal of R, the ideal *generated* by $a_1, \ldots, a_m$. An ideal of this type is said to be *finitely generated*.

We now show that if S is an ideal of the ring R, then the set $\mathscr{S}$ of all cosets $S + a$ of S can be given the structure of a ring. The ring R is a commutative group under addition. Hence, as we saw in §7, $\mathscr{S}$ acquires the structure of a (commutative) group under addition if we define the sum of $S+a$ and $S+b$ to be $S+(a+b)$. If $x = s+a$ and $x' = s' + b$ for some $s, s' \in S$, then $xx' = s'' + ab$, where $s'' = ss' + as' + sb$. Since S is an ideal, $s'' \in S$. Thus without ambiguity we may define the product of the cosets $S + a$ and $S + b$ to be the coset $S + ab$. Evidently multiplication is associative, $S + 1$ is an identity element for multiplication and both distributive laws hold. The new ring thus constructed is called the *quotient ring* of R by the ideal S, and is denoted by R/S.

A mapping $f : R \to R'$ of a ring R into a ring R' is said to be a (ring) *homomorphism* if, for all $a, b \in R$,

$$f(a + b) = f(a) + f(b), \quad f(ab) = f(a)f(b),$$

and if $f(1) = 1'$ is the identity element for multiplication in R'.

The *kernel* of the homomorphism f is the set N of all $a \in R$ such that $f(a) = 0'$ is the identity element for addition in R'. The kernel is an ideal of R, since it is a subgroup under addition and since $a \in N$, $c \in R$ imply $ac \in N$ and $ca \in N$.

For any $a \in R$, put $a' = f(a) \in R'$. The coset $N + a$ is the set of all $x \in R$ such that $f(x) = a'$, and the map $N + a \to a'$ is a bijection from the collection of all cosets of N to $f(R)$. Since f is a homomorphism, $N + (a + b)$ is mapped to $a' + b'$ and $N + ab$ is mapped to $a'b'$. Hence the map $N + a \to a'$ is also a homomorphism of the quotient ring R/N into $f(R)$.

A mapping $f : R \to R'$ of a ring R into a ring R' is said to be a (ring) *isomorphism* if it is both bijective and a homomorphism. The inverse mapping $f^{-1}: R' \to R$ is then also an isomorphism. (An *automorphism* of a ring R is an isomorphism of R with itself.)

Thus we have shown that, if $f : R \to R'$ is a homomorphism of a ring R into a ring R', with kernel N, then the quotient ring R/N is isomorphic to $f(R)$.

An ideal M of a ring R is said to be *maximal* if $M \neq R$ and if there are no ideals S such that $M \subset S \subset R$.

Let M be an ideal of the ring R. If S is an ideal of R which contains M, then the set S' of all cosets $M + a$ with $a \in S$ is an ideal of R/M. Conversely, if S' is an ideal of R/M, then the set S of all $a \in R$ such that $M + a \in S'$ is an ideal of R which contains M. It follows that M is a maximal ideal of R if and only if R/M is simple. Hence an ideal M of a commutative ring R is maximal if and only if the quotient ring R/M is a field.

To conclude, we mention a simple way of creating new rings from given ones. Let R, R' be rings and let $R \times R'$ be the set of all ordered pairs (a, a') with $a \in R$ and $a' \in R'$. As we saw in the previous section, $R \times R'$ acquires the structure of a (commutative) group under addition if we define the sum $(a, a') + (b, b')$ of (a, a') and (b, b') to be $(a + b, a' + b')$. If we define their product $(a, a') \cdot (b, b')$ to be $(ab, a'b')$, then $R \times R'$ becomes a ring, with $(0, 0')$ as identity element for addition and $(1, 1')$ as identity element for multiplication. The ring thus constructed is called the *direct sum* of R and R', and is denoted by $R \oplus R'$.

9 Vector Spaces and Associative Algebras

Although we assume some knowledge of linear algebra, it may be useful to place the basic definitions and results in the context of the preceding sections. A set V is said to be a *vector space* over a division ring D if it is a commutative group under an operation $+$ (addition) and there exists a map $\varphi : D \times V \to V$ (multiplication by a scalar) such that, if $\varphi(\alpha, v)$ is denoted by αv then, for all $\alpha, \beta \in D$ and all $v, w \in V$,

(i) $\alpha(v + w) = \alpha v + \alpha w$,
(ii) $(\alpha + \beta)v = \alpha v + \beta v$,
(iii) $(\alpha\beta)v = \alpha(\beta v)$,
(iv) $1v = v$,

where 1 is the identity element for multiplication in D. The elements of V will be called *vectors* and the elements of D *scalars*.

For example, for any positive integer n, the set D^n of all n-tuples of elements of the division ring D is a vector space over D if addition and multiplication by a scalar are defined by

$$(\alpha_1, \ldots, \alpha_n) + (\beta_1, \ldots, \beta_n) = (\alpha_1 + \beta_1, \ldots, \alpha_n + \beta_n),$$
$$\alpha(\alpha_1, \ldots, \alpha_n) = (\alpha\alpha_1, \ldots, \alpha\alpha_n).$$

The special cases $D = \mathbb{R}$ and $D = \mathbb{C}$ have many applications.

As another example, the set $\mathscr{C}(I)$ of all continuous functions $f : I \to \mathbb{R}$, where I is an interval of the real line, is a vector space over the field $\mathbb{R}$ of real numbers if addition and multiplication by a scalar are defined, for every $t \in I$, by

$$(f + g)(t) = f(t) + g(t),$$
$$(\alpha f)(t) = \alpha f(t).$$

Let V be an arbitrary vector space over a division ring D. If O is the identity element of V with respect to addition, then

$$\alpha O = O \quad \text{for every } \alpha \in D,$$

since $\alpha O = \alpha(O + O) = \alpha O + \alpha O$. Similarly, if 0 is the identity element of D with respect to addition, then

$$0v = O \quad \text{for every } v \in V,$$

since $0v = (0 + 0)v = 0v + 0v$. Furthermore,

$$(-\alpha)v = -(\alpha v) \quad \text{for all } \alpha \in D \text{ and } v \in V,$$

since $O = 0v = (\alpha + (-\alpha))v = \alpha v + (-\alpha)v$, and

$$\alpha v \neq O \quad \text{if } \alpha \neq 0 \text{ and } v \neq O,$$

since $\alpha^{-1}(\alpha v) = (\alpha^{-1}\alpha)v = 1v = v$.

From now on we will denote the zero elements of D and V by the same symbol 0. This is easier on the eye and in practice is not confusing.

A subset U of a vector space V is said to be a *subspace* of V if it is a vector space under the same operations as V itself. It is easily seen that a nonempty subset U is a subspace of V if (and only if) it is closed under addition and multiplication by a scalar. For then, if $u \in U$, also $-u = (-1)u \in U$, and so U is an additive subgroup of V. The other requirements for a vector space are simply inherited from V.

For example, if $1 \leq m < n$, the set of all $(\alpha_1, \ldots, \alpha_n) \in D^n$ with $\alpha_1 = \cdots = \alpha_m = 0$ is a subspace of D^n. Also, the set $\mathscr{C}^1(I)$ of all continuously differentiable functions $f : I \to \mathbb{R}$ is a subspace of $\mathscr{C}(I)$. Two obvious subspaces of any vector space V are V itself and the subset $\{0\}$ which contains only the zero vector.

If U_1 and U_2 are subspaces of a vector space V, then their *intersection* $U_1 \cap U_2$, which necessarily contains 0, is again a subspace of V. The *sum* $U_1 + U_2$, consisting of all vectors $u_1 + u_2$ with $u_1 \in U_1$ and $u_2 \in U_2$, is also a subspace of V. Evidently $U_1 + U_2$ contains U_1 and U_2 and is contained in every subspace of V which contains both U_1 and U_2. If $U_1 \cap U_2 = \{0\}$, the sum $U_1 + U_2$ is said to be *direct*, and is denoted by $U_1 \oplus U_2$, since it may be identified with the set of all ordered pairs (u_1, u_2), where $u_1 \in U_1$ and $u_2 \in U_2$.

Let V be an arbitrary vector space over a division ring D and let $\{v_1, \ldots, v_m\}$ be a finite subset of V. A vector v in V is said to be a *linear combination* of $v_1, \ldots, v_m$ if

$$v = \alpha_1 v_1 + \cdots + \alpha_m v_m$$

for some $\alpha_1, \ldots, \alpha_m \in D$. The *coefficients* $\alpha_1, \ldots, \alpha_m$ need not be uniquely determined. Evidently a vector v is a linear combination of $v_1, \ldots, v_m$ if it is a linear combination of some proper subset, since we can add the remaining vectors with zero coefficients.

If S is any nonempty subset of V, then the set $\langle S \rangle$ of all vectors in V which are linear combinations of finitely many elements of S is a subspace of V, the subspace 'spanned' or *generated* by S. Clearly $S \subseteq \langle S \rangle$ and $\langle S \rangle$ is contained in every subspace of V which contains S.

A finite subset $\{v_1, \ldots, v_m\}$ of V is said to be *linearly dependent* (over D) if there exist $\alpha_1, \ldots, \alpha_m \in D$, not all zero, such that

$$\alpha_1 v_1 + \cdots + \alpha_m v_m = 0,$$

and is said to be *linearly independent* otherwise.

For example, in $\mathbb{R}^3$ the vectors

$$v_1 = (1, 0, 1), \quad v_2 = (1, 1, 0), \quad v_3 = (1, 1/2, 1/2)$$

are linearly dependent, since $v_1 + v_2 - 2v_3 = 0$. On the other hand, the vectors

$$e_1 = (1, 0, 0), \quad e_2 = (0, 1, 0), \quad e_3 = (0, 0, 1)$$

are linearly independent, since $\alpha_1 e_1 + \alpha_2 e_2 + \alpha_3 e_3 = (\alpha_1, \alpha_2, \alpha_3)$, and this is 0 only if $\alpha_1 = \alpha_2 = \alpha_3 = 0$.

In any vector space V, the set $\{v\}$ containing the single vector v is linearly independent if $v \neq 0$ and linearly dependent if $v = 0$. If $v_1, \ldots, v_m$ are linearly independent, then any vector $v \in \langle v_1, \ldots, v_m \rangle$ has a unique representation as a linear combination of $v_1, \ldots, v_m$, since if

$$\alpha_1 v_1 + \cdots + \alpha_m v_m = \beta_1 v_1 + \cdots + \beta_m v_m,$$

then

$$(\alpha_1 - \beta_1) v_1 + \cdots + (\alpha_m - \beta_m) v_m = 0$$

and hence

$$\alpha_1 - \beta_1 = \cdots = \alpha_m - \beta_m = 0.$$

Evidently the vectors $v_1, \ldots, v_m$ are linearly dependent if some proper subset is linearly dependent. Hence any nonempty subset of a linearly independent set is again linearly independent.

A subset S of a vector space V is said to be a *basis* for V if S is linearly independent and $\langle S \rangle = V$. In the previous example, the vectors e_1, e_2, e_3 are a basis for $\mathbb{R}^3$, since they are not only linearly independent but also generate $\mathbb{R}^3$.

Any nontrivial finitely generated vector space has a basis. In fact if a vector space V is generated by a finite subset T, then V has a basis $B \subseteq T$. Moreover, any linearly independent subset of V is also finite and its cardinality does not exceed that of T. It follows that any two bases contain the same number of elements.

If V has a basis containing n elements, we say V has *dimension* n and we write $\dim V = n$. We say that V has infinite dimension if it is not finitely generated, and has dimension 0 if it contains only the vector 0.

For example, the field $\mathbb{C}$ of complex numbers may be regarded as a 2-dimensional vector space over the field $\mathbb{R}$ of real numbers, with basis $\{1, i\}$.

Again, D^n has dimension n as a vector space over the division ring D, since it has the basis

$$e_1 = (1, 0, \ldots, 0), \quad e_2 = (0, 1, \ldots, 0), \ldots, \quad e_n = (0, 0, \ldots, 1).$$

On the other hand, the real vector space $\mathscr{C}(I)$ of all continuous functions $f : I \to \mathbb{R}$ has infinite dimension if the interval I contains more than one point since, for any positive integer n, the real polynomials of degree less than n form an n-dimensional subspace.

The first of these examples is readily generalized. If E and F are fields with $F \subseteq E$, we can regard E as a vector space over F. If this vector space is finite-dimensional, we say that E is a *finite extension* of F and define the *degree* of E over F to be the dimension $[E : F]$ of this vector space.

Any subspace U of a finite-dimensional vector space V is again finite-dimensional. Moreover, $\dim U \leq \dim V$, with equality only if $U = V$. If U_1 and U_2 are subspaces of V, then

$$\dim(U_1 + U_2) + \dim(U_1 \cap U_2) = \dim U_1 + \dim U_2.$$

Let V and W be vector spaces over the same division ring D. A map $T : V \to W$ is said to be *linear*, or a *linear transformation*, or a 'vector space homomorphism', if for all $v, v' \in V$ and every $\alpha \in D$,

$$T(v + v') = Tv + Tv', \quad T(\alpha v) = \alpha(Tv).$$

Since the first condition implies that T is a homomorphism of the additive group of V into the additive group of W, it follows that $T0 = 0$ and $T(-v) = -Tv$.

For example, if (τ_{jk}) is an $m \times n$ matrix with entries from the division ring D, then the map $T : D^m \to D^n$ defined by

$$T(\alpha_1, \ldots, \alpha_m) = (\beta_1, \ldots, \beta_n),$$

where

$$\beta_k = \alpha_1 \tau_{1k} + \cdots + \alpha_m \tau_{mk} \quad (1 \leq k \leq n),$$

is linear. It is easily seen that every linear map of D^m into D^n may be obtained in this way.

As another example, if $\mathscr{C}^1(I)$ is the real vector space of all continuously differentiable functions $f : I \to R$, then the map $T : \mathscr{C}^1(I) \to \mathscr{C}(I)$ defined by $Tf = f'$ (the derivative of f) is linear.

Let U, V, W be vector spaces over the same division ring D. If $T : V \to W$ and $S : U \to V$ are linear maps, then the composite map $T \circ S : U \to W$ is again linear. For linear maps it is customary to write TS instead of $T \circ S$. The identity map

$I : V \to V$ defined by $Iv = v$ for every $v \in V$ is clearly linear. If a linear map $T : V \to W$ is bijective, then its inverse map $T^{-1} : W \to V$ is again linear.

If $T : V \to W$ is a linear map, then the set N of all $v \in V$ such that $Tv = 0$ is a subspace of V, called the *nullspace* or *kernel* of T. Since $Tv = Tv'$ if and only if $T(v - v') = 0$, the map T is injective if and only if its kernel is $\{0\}$, i.e. when T is *nonsingular*.

For any subspace U of V, its image $TU = \{Tv : v \in U\}$ is a subspace of W. In particular, TV is a subspace of W, called the *range* of T. Thus the map T is surjective if and only if its range is W.

If V is finite-dimensional, then the range R of T is also finite-dimensional and

$$\dim R = \dim V - \dim N,$$

(since $R \approx V/N$). The dimensions of R and N are called respectively the *rank* and *nullity* of T. It follows that, if $\dim V = \dim W$, then T is injective if and only if it is surjective.

Two vector spaces V, W over the same division ring D are said to be *isomorphic* if there exists a bijective linear map $T : V \to W$. As an example, if V is an n-dimensional vector space over the division ring D, then V is isomorphic to D^n. For if $v_1, \ldots, v_n$ is a basis for V and if $v = \alpha_1 v_1 + \cdots + \alpha_n v_n$ is an arbitrary element of V, the map $v \to (\alpha_1, \ldots, \alpha_n)$ is linear and bijective.

Thus there is essentially only one vector space of given finite dimension over a given division ring. However, vector spaces do not always present themselves in the concrete form D^n. An example is the set of solutions of a system of homogeneous linear equations with real coefficients. Hence, even if one is only interested in the finite-dimensional case, it is still desirable to be acquainted with the abstract definition of a vector space.

Let V and W be vector spaces over the same division ring D. We can define the *sum* $S + T$ of two linear maps $S : V \to W$ and $T : V \to W$ by

$$(S + T)v = Sv + Tv.$$

This is again a linear map, and it is easily seen that with this definition of addition the set of all linear maps of V into W is a commutative group. If D is a field, i.e. if multiplication in D is commutative, then for any $\alpha \in D$ the map αT defined by

$$(\alpha T)v = \alpha(Tv)$$

is again linear, and with these definitions of addition and multiplication by a scalar the set of all linear maps of V into W is a vector space over D. (If the division ring D is not a field, it is necessary to consider 'right' vector spaces over D, as well as 'left' ones.)

If $V = W$, then the *product* TS is also defined and it is easily verified that the set of all linear maps of V into itself is a ring, with the identity map I as identity element for multiplication. The bijective linear maps of V to itself are the units of this ring and thus form a group under multiplication, the *general linear group* $GL(V)$.

Similarly to the direct product of two groups and the direct sum of two rings, one may define the *tensor product* $V \otimes V'$ of two vector spaces V, V' and the *Kronecker product* $T \otimes T'$ of two linear maps $T : V \to W$ and $T' : V' \to W'$.

The *centre* of a ring R is the set of all $c \in R$ such that $ac = ca$ for every $a \in R$. An *associative algebra* A over a field F is a ring containing F in its centre. On account of

the ring structure, we can regard A as a vector space over F. The associative algebra is said to be *finite-dimensional* if it is finite-dimensional as a vector space over F.

For example, the set $M_n(F)$ of all $n \times n$ matrices with entries from the field F is a finite-dimensional associative algebra, with the usual definitions of addition and multiplication, and with $\alpha \in F$ identified with the matrix αI.

More generally, if D is a division ring containing F in its centre, then the set $M_n(D)$ of all $n \times n$ matrices with entries from D is an associative algebra over F. It is finite-dimensional if D itself is finite dimensional over F.

By the definition for rings, an associative algebra A is *simple* if $A \neq \{0\}$ and A has no ideals except $\{0\}$ and A. It is not difficult to show that, for any division ring D containing F in its centre, the associative algebra $M_n(D)$ is simple. It was proved by Wedderburn (1908) that any finite-dimensional simple associative algebra has the form $M_n(D)$, where D is a division ring containing F in its centre and of finite dimension over F.

If $F = \mathbb{C}$, the fundamental theorem of algebra implies that $\mathbb{C}$ is the only such D. If $F = \mathbb{R}$, there are three choices for D, by the following theorem of Frobenius (1878):

Proposition 31 *If a division ring D contains the real field $\mathbb{R}$ in its centre and is of finite dimension as a vector space over $\mathbb{R}$, then D is isomorphic to $\mathbb{R}$, $\mathbb{C}$ or $\mathbb{H}$.*

Proof Suppose first that D is a field and $D \neq \mathbb{R}$. If $a \in D \backslash \mathbb{R}$ then, since D is finite-dimensional over $\mathbb{R}$, a is a root of a monic polynomial with real coefficients, which we may assume to be of minimal degree. Since $a \notin \mathbb{R}$, the degree is not 1 and the fundamental theorem of algebra implies that it must be 2. Thus

$$a^2 - 2\lambda a + \mu = 0$$

for some $\lambda, \mu \in \mathbb{R}$ with $\lambda^2 < \mu$. Then $\mu - \lambda^2 = \rho^2$ for some nonzero $\rho \in \mathbb{R}$ and $i = (a - \lambda)/\rho$ satisfies $i^2 = -1$. Thus D contains the field $\mathbb{R}(i) = \mathbb{R} + i\mathbb{R}$. But, since D is a field, the only $x \in D$ such that $x^2 = -1$ are i and $-i$. Hence the preceding argument shows that actually $D = \mathbb{R}(i)$. Thus D is isomorphic to the field $\mathbb{C}$ of complex numbers.

Suppose now that D is not commutative. Let a be an element of D which is not in the centre of D, and let M be an $\mathbb{R}$-subspace of D of maximal dimension which is commutative and which contains both a and the centre of D. If $x \in D$ commutes with every element of M, then $x \in M$. Hence M is a maximal commutative subset of D. It follows that if $x \in M$ and $x \neq 0$ then also $x^{-1} \in M$, since $xy = yx$ for all $y \in M$ implies $yx^{-1} = x^{-1}y$ for all $y \in M$. Similarly $x, x' \in M$ implies $xx' \in M$. Thus M is a field which properly contains $\mathbb{R}$. Hence, by the first part of the proof, M is isomorphic to $\mathbb{C}$. Thus $M = \mathbb{R}(i)$, where $i^2 = -1$, $[M : \mathbb{R}] = 2$ and $\mathbb{R}$ is the centre of D.

If $x \in D \backslash M$, then $b = (x + ixi)/2$ satisfies

$$bi = (xi - ix)/2 = -ib \neq 0.$$

Hence $b \in D \backslash M$ and $b^2 i = ib^2$. But, in the same way as before, $N = \mathbb{R} + \mathbb{R}b$ is a maximal subfield of D containing b and $\mathbb{R}$, and $N = \mathbb{R}(j)$, where $j^2 = -1$. Thus $b^2 = \alpha + \beta b$, where $\alpha, \beta \in \mathbb{R}$. In fact, since $b^2 i = ib^2$, we must have $\beta = 0$. Similarly

$j = \gamma + \delta b$, where $\gamma, \delta \in \mathbb{R}$ and $\delta \neq 0$. Since $j^2 = \gamma^2 + 2\gamma\delta b + \delta^2 \alpha = -1$, we must have $\gamma = 0$. Thus $j = \delta b$ and $ji = -ij$.

If we put $k = ij$, it now follows that

$$k^2 = -1, \quad jk = i = -kj, \quad ki = j = -ik.$$

Since no $\mathbb{R}$-linear combination of $1, i, j$ has these properties, the elements $1, i, j, k$ are $\mathbb{R}$-linearly independent. But, by Proposition 32 below, $[D : M] = [M : \mathbb{R}] = 2$. Hence $[D : \mathbb{R}] = 4$ and $1, i, j, k$ are a basis for D over $\mathbb{R}$. Thus D is isomorphic to the division ring $\mathbb{H}$ of quaternions. □

To complete the proof of Proposition 31 we now prove

Proposition 32 *Let D be a division ring which, as a vector space over its centre C, has finite dimension $[D : C]$. If M is a maximal subfield of D, then $[D : M] = [M : C]$.*

Proof Put $n = [D : C]$ and let $e_1, \ldots, e_n$ be a basis for D as a vector space over C. Obviously we may suppose $n > 1$. We show first that if $a_1, \ldots, a_n$ are elements of D such that

$$a_1 x e_1 + \cdots + a_n x e_n = 0 \quad \text{for every } x \in D,$$

then $a_1 = \cdots = a_n = 0$. Assume that there exists such a set $\{a_1, \ldots, a_n\}$ with not all elements zero and choose one with the minimal number of nonzero elements. We may suppose the notation chosen so that $a_i \neq 0$ for $i \leq r$ and $a_i = 0$ for $i > r$ and, by multiplying on the left by a_1^{-1}, we may further suppose that $a_1 = 1$. For any $y \in D$ we have

$$a_1 y x e_1 + \cdots + a_n y x e_n = 0 = y(a_1 x e_1 + \cdots + a_n x e_n)$$

and hence

$$(a_1 y - y a_1) x e_1 + \cdots + (a_n y - y a_n) x e_n = 0.$$

Since $a_i y = y a_i$ for $i = 1$ and for $i > r$, our choice of $\{a_1, \ldots, a_n\}$ implies that $a_i y = y a_i$ for all i. Since this holds for every $y \in D$, it follows that $a_i \in C$ for all i. But this is a contradiction, since $e_1, \ldots, e_n$ is a basis for D over C and $a_1 e_1 + \cdots + a_n e_n = 0$.

The map $T_{jk} : D \to D$ defined by $T_{jk} x = e_j x e_k$ is a linear transformation of D as a vector space over C. By what we have just proved, the n^2 linear maps T_{jk} $(j, k = 1, \ldots, n)$ are linearly independent over C. Consequently every linear transformation of D as a vector space over C is a C-linear combination of the maps T_{jk}.

Suppose now that $T : D \to D$ is a linear transformation of D as a vector space over M. Since $C \subseteq M$, T is also a linear transformation of D as a vector space over C and hence has the form

$$Tx = a_1 x e_1 + \cdots + a_n x e_n$$

for some $a_1, \ldots, a_n \in D$. But $T(bx) = b(Tx)$ for all $b \in M$ and $x \in D$. Hence

$$(a_1 b - b a_1) x e_1 + \cdots + (a_n b - b a_n) x e_n = 0 \quad \text{for every } x \in D,$$

which implies $a_i b = b a_i$ $(i = 1, \ldots, n)$. Since this holds for all $b \in M$ and M is a maximal subfield of D, it follows that $a_i \in M$ $(i = 1, \ldots, n)$.

Let $\mathscr{T}$ denote the set of all linear transformations of D as a vector space over M. By what we have already proved, every $T \in \mathscr{T}$ is an M-linear combination of the maps $T_1, \ldots, T_n$, where $T_i x = x e_i$ $(i = 1, \ldots, n)$, and the maps $T_1, \ldots, T_n$ are linearly independent over M. Consequently the dimension of $\mathscr{T}$ as a vector space over M is n. But $\mathscr{T}$ has dimension $[D : M]^2$ as a vector space over M. Hence $[D : M]^2 = n$. Since $n = [D : M][M : C]$, it follows that $[D : M] = [M : C]$. □

10 Inner Product Spaces

Let F denote either the real field $\mathbb{R}$ or the complex field $\mathbb{C}$. A vector space V over F is said to be an *inner product space* if there exists a map $(u, v) \to \langle u, v \rangle$ of $V \times V$ into F such that for every $\alpha \in F$ and all $u, u', v \in V$,

(i) $\langle \alpha u, v \rangle = \alpha \langle u, v \rangle$,
(ii) $\langle u + u', v \rangle = \langle u, v \rangle + \langle u', v \rangle$,
(iii) $\langle v, u \rangle = \overline{\langle u, v \rangle}$,
(iv) $\langle u, u \rangle > 0$ if $u \neq O$.

If $F = \mathbb{R}$, then (iii) simply says that $\langle v, u \rangle = \langle u, v \rangle$, since a real number is its own complex conjugate. The restriction $u \neq O$ is neeessary in (iv), since (i) and (iii) imply that

$$\langle u, O \rangle = \langle O, v \rangle = 0 \quad \text{for all } u, v \in V.$$

It follows from (ii) and (iii) that

$$\langle u, v + v' \rangle = \langle u, v \rangle + \langle u, v' \rangle \quad \text{for all } u, v, v' \in V,$$

and from (i) and (iii) that

$$\langle u, \alpha v \rangle = \bar{\alpha} \langle u, v \rangle \quad \text{for every } \alpha \in F \text{ and all } u, v \in V.$$

The standard example of an inner product space is the vector space F^n, with the inner product of $x = (\xi_1, \ldots, \xi_n)$ and $y = (\eta_1, \ldots, \eta_n)$ defined by

$$\langle x, y \rangle = \xi_1 \bar{\eta}_1 + \cdots + \xi_n \bar{\eta}_n.$$

Another example is the vector space $\mathscr{C}(I)$ of all continuous functions $f : I \to F$, where $I = [a, b]$ is a compact subinterval of $\mathbb{R}$, with the inner product of f and g defined by

$$\langle f, g \rangle = \int_a^b f(t) \overline{g(t)} dt.$$

In an arbitrary inner product space V we define the *norm* $\|v\|$ of a vector $v \in V$ by

$$\|v\| = \langle v, v \rangle^{1/2}.$$

Thus $\|v\| \geq 0$, with equality if and only if $v = O$. Evidently

$$\|\alpha v\| = |\alpha|\|v\| \quad \text{for all } \alpha \in F \text{ and } v \in V.$$

Inner products and norms are connected by *Schwarz's inequality*:

$$|\langle u, v\rangle| \leq \|u\|\|v\| \quad \text{for all } u, v \in V,$$

with equality if and only if u and v are linearly dependent. For the proof we may suppose that u and v are linearly independent, since it is easily seen that equality holds if $u = \lambda v$ or $v = \lambda u$ for some $\lambda \in F$. Then, for all $\alpha, \beta \in F$, not both 0,

$$0 < \langle \alpha u + \beta v, \alpha u + \beta v\rangle = |\alpha|^2\langle u, u\rangle + \alpha\bar{\beta}\langle u, v\rangle + \bar{\alpha}\beta\overline{\langle u, v\rangle} + |\beta|^2\langle v, v\rangle.$$

If we choose $\alpha = \langle v, v\rangle$ and $\beta = -\langle u, v\rangle$, this takes the form

$$0 < \|u\|^2\|v\|^4 - 2\|v\|^2|\langle u, v\rangle|^2 + |\langle u, v\rangle|^2\|v\|^2 = \{\|u\|^2\|v\|^2 - |\langle u, v\rangle|^2\}\|v\|^2.$$

Hence

$$|\langle u, v\rangle|^2 < \|u\|^2\|v\|^2,$$

as we wished to show. We follow common practice by naming the inequality after Schwarz (1885), but (cf. §4) it had already been proved for $\mathbb{R}^n$ by Cauchy (1821) and for $\mathscr{C}(I)$ by Bunyakovskii (1859).

It follows from Schwarz's inequality that

$$\begin{aligned}\|u + v\|^2 &= \|u\|^2 + 2\mathscr{R}\langle u, v\rangle + \|v\|^2 \\ &\leq \|u\|^2 + 2|\langle u, v\rangle| + \|v\|^2 \leq \{\|u\| + \|v\|\}^2.\end{aligned}$$

Thus

$$\|u + v\| \leq \|u\| + \|v\| \quad \text{for all } u, v \in V,$$

with strict inequality if u and v are linearly independent.

It now follows that V acquires the structure of a metric space if we define the distance between u and v by

$$\mathrm{d}(u, v) = \|u - v\|.$$

In the case $V = \mathbb{R}^n$ this is the *Euclidean distance*

$$\mathrm{d}(x, y) = \left(\sum_{j=1}^{n} |\xi_j - \eta_j|^2\right)^{1/2},$$

and in the case $V = \mathscr{C}(I)$ it is the *L^2-norm*

$$\mathrm{d}(f, g) = \left(\int_a^b |f(t) - g(t)|^2 dt\right)^{1/2}.$$

The norm in any inner product space V satisfies the *parallelogram law*:

$$\|u+v\|^2+\|u-v\|^2=2\|u\|^2+2\|v\|^2 \quad \text{for all } u,v\in V.$$

This may be immediately verified by substituting $\|w\|^2=\langle w,w\rangle$ throughout and using the linearity of the inner product. The geometrical interpretation is that in any parallelogram the sum of the squares of the lengths of the two diagonals is equal to the sum of the squares of the lengths of all four sides.

It may be shown that any normed vector space which satisfies the parallelogram law can be given the structure of an inner product space by defining

$$\begin{aligned}\langle u,v\rangle &= \{\|u+v\|^2-\|u-v\|^2\}/4 \quad \text{if } F=\mathbb{R},\\ &= \{\|u+v\|^2-\|u-v\|^2+i\|u+iv\|^2-i\|u-iv\|^2\}/4 \quad \text{if } F=\mathbb{C}.\end{aligned}$$

(Cf. the argument for $F=\mathbb{Q}$ in §4 of Chapter XIII.)

In an arbitrary inner product space V a vector u is said to be 'perpendicular' or *orthogonal* to a vector v if $\langle u,v\rangle=0$. The relation is symmetric, since $\langle u,v\rangle=0$ implies $\langle v,u\rangle=0$. For orthogonal vectors u,v, the *law of Pythagoras* holds:

$$\|u+v\|^2=\|u\|^2+\|v\|^2.$$

More generally, a subset E of V is said to be *orthogonal* if $\langle u,v\rangle=0$ for all $u,v\in E$ with $u\neq v$. It is said to be *orthonormal* if, in addition, $\langle u,u\rangle=1$ for every $u\in E$. An orthogonal set which does not contain O may be converted into an orthonormal set by replacing each $u\in E$ by $u/\|u\|$.

For example, if $V=F^n$, then the basis vectors

$$e_1=(1,0,\ldots,0),\quad e_2=(0,1,\ldots,0),\ldots,\quad e_n=(0,0,\ldots,1)$$

form an orthonormal set. It is easily verified also that, if $I=[0,1]$, then in $\mathscr{C}(I)$ the functions $e_n(t)=e^{2\pi int}$ $(n\in\mathbb{Z})$ form an orthonormal set.

Let $\{e_1,\ldots,e_m\}$ be *any* orthonormal set in the inner product space V and let U be the vector subspace generated by $e_1,\ldots,e_m$. Then the norm of a vector $u=\alpha_1e_1+\cdots+\alpha_me_m\in U$ is given by

$$\|u\|^2=|\alpha_1|^2+\cdots+|\alpha_m|^2,$$

which shows that $e_1,\ldots,e_m$ are linearly independent.

To find the *best approximation* in U to a given vector $v\in V$, put

$$w=\gamma_1e_1+\cdots+\gamma_me_m,$$

where

$$\gamma_j=\langle v,e_j\rangle \quad (j=1,\ldots,m).$$

Then $\langle w,e_j\rangle=\langle v,e_j\rangle (j=1,\ldots,m)$ and hence $\langle v-w,w\rangle=0$. Consequently, by the law of Pythagoras,

$$\|v\|^2=\|v-w\|^2+\|w\|^2.$$

Since $\|w\|^2 = |\gamma_1|^2 + \cdots + |\gamma_m|^2$, this yields *Bessel's inequality*:

$$|\langle v, e_1\rangle|^2 + \cdots + |\langle v, e_m\rangle|^2 \leq \|v\|^2,$$

with strict inequality if $v \notin U$. For any $u \in U$, we also have $\langle v - w, w - u\rangle = 0$ and so, by Pythagoras again,

$$\|v - u\|^2 = \|v - w\|^2 + \|w - u\|^2.$$

This shows that w is the unique nearest point of U to v.

From any linearly independent set of vectors $v_1, \ldots, v_m$ we can inductively construct an orthonormal set $e_1, \ldots, e_m$ such that $e_1, \ldots, e_k$ span the same vector subspace as $v_1, \ldots, v_k$ for $1 \leq k \leq m$. We begin by taking $e_1 = v_1/\|v_1\|$. Now suppose $e_1, \ldots, e_k$ have been determined. If

$$w = v_{k+1} - \langle v_{k+1}, e_1\rangle e_1 - \cdots - \langle v_{k+1}, e_k\rangle e_k,$$

then $\langle w, e_j\rangle = 0$ $(j = 1, \ldots, k)$. Moreover $w \neq O$, since w is a linear combination of $v_1, \ldots, v_{k+1}$ in which the coefficient of v_{k+1} is 1. By taking $e_{k+1} = w/\|w\|$, we obtain an orthonormal set $e_1, \ldots, e_{k+1}$ spanning the same linear subspace as $v_1, \ldots, v_{k+1}$. This construction is known as *Schmidt's orthogonalization process*, because of its use by E. Schmidt (1907) in his treatment of linear integral equations. The (normalized) Legendre polynomials are obtained by applying the process to the linearly independent functions $1, t, t^2, \ldots$ in the space $\mathscr{C}(I)$, where $I = [-1, 1]$.

It follows that any finite-dimensional inner product space V has an orthonormal basis $e_1, \ldots, e_n$ and that

$$\|v\|^2 = \sum_{j=1}^{n} |\langle v, e_j\rangle|^2 \quad \text{for every } v \in V.$$

In an infinite-dimensional inner product space V an orthonormal set E may even be uncountably infinite. However, for a given $v \in V$, there are at most countably many vectors $e \in E$ for which $\langle v, e\rangle \neq 0$. For if $\{e_1, \ldots, e_m\}$ is any finite subset of E then, by Bessel's inequality,

$$\sum_{j=1}^{m} |\langle v, e_j\rangle|^2 \leq \|v\|^2$$

and so, for each $n \in \mathbb{N}$, there are at most $n^2 - 1$ vectors $e \in E$ for which $|\langle v, e\rangle| > \|v\|/n$.

If the vector subspace U of all finite linear combinations of elements of E is dense in V then, by the best approximation property of finite orthonormal sets, *Parseval's equality* holds:

$$\sum_{e \in E} |\langle v, e\rangle|^2 = \|v\|^2 \quad \text{for every } v \in V.$$

Parseval's equality holds for the inner product space $\mathscr{C}(I)$, where $I = [0, 1]$, and the orthonormal set $E = \{e^{2\pi i n t} : n \in \mathbb{Z}\}$ since, by *Weierstrass's approximation theorem* (see the references in §6 of Chapter XI), every $f \in \mathscr{C}(I)$ is the uniform limit of a sequence of *trigonometric polynomials*. The result in this case was formally derived by Parseval (1805).

An *almost periodic function*, in the sense of Bohr (1925), is a function $f : \mathbb{R} \to \mathbb{C}$ which can be uniformly approximated on $\mathbb{R}$ by *generalized trigonometric polynomials*

$$\sum_{j=1}^{m} c_j e^{i\lambda_j t},$$

where $c_j \in \mathbb{C}$ and $\lambda_j \in \mathbb{R}$ $(j = 1, \ldots, m)$. For any almost periodic functions f, g, the limit

$$\langle f, g \rangle = \lim_{T\to\infty} (1/2T) \int_{-T}^{T} f(t)\overline{g(t)}dt$$

exists. The set $\mathscr{B}$ of all almost periodic functions acquires in this way the structure of an inner product space. The set $E = \{e^{i\lambda t} : \lambda \in \mathbb{R}\}$ is an uncountable orthonormal set and Parseval's equality holds for this set.

A finite-dimensional inner product space is necessarily complete as a metric space, i.e., every fundamental sequence converges. However, an infinite-dimensional inner product space need not be complete, as $\mathscr{C}(I)$ already illustrates. An inner product space which is complete is said to be a *Hilbert space*.

The case considered by Hilbert (1906) was the vector space ℓ^2 of all infinite sequences $x = (\xi_1, \xi_2, \ldots)$ of complex numbers such that $\sum_{k\geq 1} |\xi_k|^2 < \infty$, with

$$\langle x, y \rangle = \sum_{k\geq 1} \xi_k \bar{\eta}_k.$$

Another example is the vector space $L^2(I)$, where $I = [0, 1]$, of all (equivalence classes of) Lebesgue measurable functions $f : I \to \mathbb{C}$ such that $\int_0^1 |f(t)^2 dt < \infty$, with

$$\langle f, g \rangle = \int_0^1 f(t)\overline{g(t)}dt.$$

With any $f \in L^2(I)$ we can associate a sequence $\hat{f} \in \ell^2$, consisting of the inner products $\langle f, e_n \rangle$, where $e_n(t) = e^{2\pi int}$ $(n \in \mathbb{Z})$, in some fixed order. The map $\mathscr{F} : L^2(I) \to \ell^2$ thus defined is linear and, by Parseval's equality,

$$\|\mathscr{F}f\| = \|f\|.$$

In fact $\mathscr{F}$ is an *isometry* since, by the *theorem of Riesz–Fischer* (1907), it is bijective.

11 Further Remarks

A vast fund of information about numbers in different cultures is contained in Menninger [52]. A good popular book is Dantzig [18].

The algebra of sets was created by Boole (1847), who used the symbols $+$ and $\cdot$ instead of $\cup$ and $\cap$, as is now customary. His ideas were further developed, with applications to logic and probability theory, in Boole [10]. A simple system of axioms for

Boolean algebra was given by Huntingdon [39]. For an introduction to Stone's representation theorem, referred to in §8, see Stone [69]; there are proofs in Halmos [30] and Sikorski [66]. For applications of Boolean algebras to switching circuits see, for example, Rudeanu [62]. Boolean algebra is studied in the more general context of lattice theory in Birkhoff [6].

Dedekind's axioms for $\mathbb{N}$ may be found on p. 67 of [19], which contains also his earlier construction of the real numbers from the rationals by means of cuts. Some interesting comments on the axioms **(N1)–(N3)** are contained in Henkin [34]. Starting from these axioms, Landau [47] gives a detailed derivation of the basic properties of $\mathbb{N}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$.

The argument used to extend $\mathbb{N}$ to $\mathbb{Z}$ shows that any commutative *semigroup* satisfying the cancellation law may be embedded in a commutative *group*. The argument used to extend $\mathbb{Z}$ to $\mathbb{Q}$ shows that any commutative *ring* without divisors of zero may be embedded in a *field*.

An example of an ordered field which does not have the Archimedean property, although every fundamental sequence is (trivially) convergent, is the field $*\mathbb{R}$ of hyperreal numbers, constructed by Abraham Robinson (1961). Hyperreal numbers are studied in Stroyan and Luxemburg [70].

The 'arithmetization of analysis' had a gradual evolution, which is traced in Chapitre VI (by Dugac) of Dieudonné *et al*. [22]. A modern text on real analysis is Rudin [63]. In Lemma 7 of Chapter VI we will show that all norms on $\mathbb{R}^n$ are equivalent.

The contraction principle (Proposition 26) has been used to prove the *central limit theorem* of probability theory by Hamedani and Walter [32]. Bessaga (1959) has proved a *converse* of the contraction principle: Let E be an arbitrary set, $f : E \to E$ a map of E to itself and θ a real number such that $0 < \theta < 1$. If each iterate $f^n (n \in \mathbb{N})$ has at most one fixed point and if some iterate has a fixed point, then a complete metric d can be defined on E such that $\mathrm{d}(f(x'), f(x'')) \leq \theta \mathrm{d}(x', x'')$ for all $x', x'' \in E$. A short proof is given by Jachymski [40].

There are other important fixed point theorems besides Proposition 26. *Brouwer's fixed point theorem* states that, if $B = \{x \in \mathbb{R}^n : |x| \leq 1\}$ is the n-dimensional closed unit ball, every continuous map $f : B \to B$ has a fixed point. For an elementary proof, see Kulpa [44]. The *Lefschetz fixed point theorem* requires a knowledge of algebraic topology, even for its statement. Fixed point theorems are extensively treated in Dugundji and Granas [23] (and in A. Granas and J. Dugundji, *Fixed Point Theory*, Springer-Verlag, New York, 2003).

For a more detailed discussion of differentiability for functions of several variables see, for example, Fleming [26] and Dieudonné [21]. The inverse function theorem (Proposition 27) is a local result. Some global results are given by Atkinson [5] and Chichilnisky [14]. For a holomorphic version of Proposition 28 and for the simple way in which higher-order equations may be replaced by systems of first-order equations see, e.g., Coddington and Levinson [16].

The formula for the roots of a cubic was first published by Cardano [12], but it was discovered by del Ferro and again by Tartaglia, who accused Cardano of breaking a pledge of secrecy. Cardano is judged less harshly by historians today than previously. His book, which contained developments of his own and also the formula for

the roots of a quartic discovered by his pupil Ferrari, was the most significant Western contribution to mathematics for more than a thousand years.

Proposition 29 still holds, but is more difficult to prove, if in its statement "has a nonzero derivative" is replaced by "which is not constant". Read [57] shows that the basic results of complex analysis may be deduced from this stronger form of Proposition 29 without the use of complex integration.

A field F is said to be *algebraically closed* if every polynomial of positive degree with coefficients from F has a root in F. Thus the 'fundamental theorem of algebra' says that the field $\mathbb{C}$ of complex numbers is algebraically closed. The proofs of this theorem due to Argand–Cauchy and Euler–Lagrange–Laplace are given in Chapter 4 (by Remmert) of Ebbinghaus *et al.* [24]. As shown on p. 77 of [24], the latter method provides, in particular, a simple proof for the existence of n-th roots.

Wall [72] gives a proof of the fundamental theorem of algebra, based on the notion of topological degree, and Ahlfors [1] gives the most common complex analysis proof, based on Liouville's theorem that a function holomorphic in the whole complex plane is bounded only if it is a constant. A form of Liouville's theorem is easily deduced from Proposition 29: if the power series

$$p(z) = a_0 + a_1 z + a_2 z^2 + \cdots$$

converges and $|p(z)|$ is bounded for all $z \in \mathbb{C}$, then $a_n = 0$ for every $n \geq 1$.

The representation of trigonometric functions by complex exponentials appears in §138 of Euler [25]. The various algebraic formulas involving trigonometric functions, such as

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

are easily established by means of this representation and the addition theorem for the exponential function.

Some texts on complex analysis are Ahlfors [1], Caratheodory [11] and Narasimhan [56].

The 19th century literature on quaternions is surveyed in Rothe [59]. Although Hamilton hoped that quaternions would prove as useful as complex numbers, a quaternionic analysis analogous to complex analysis was first developed by Fueter (1935). A good account is given by Sudbery [71].

One significant contribution of quaternions was indirect. After Hamilton had shown the way, other 'hypercomplex' number systems were constructed, which led eventually to the structure theory of associative algebras discussed below.

It is not difficult to show that any *automorphism* of $\mathbb{H}$, i.e. any bijective map $T : \mathbb{H} \to \mathbb{H}$ such that

$$T(x + y) = Tx + Ty, \quad T(xy) = (Tx)(Ty) \quad \text{for all } x, y \in \mathbb{H},$$

has the form $Tx = uxu^{-1}$ for some quaternion u with norm 1.

For octonions and their uses, see van der Blij [8] and Springer and Veldkamp [67]. The group of all automorphisms of the algebra $\mathbb{O}$ is the exceptional simple Lie group G_2. The other four exceptional simple Lie groups are also all related to $\mathbb{O}$ in some way.

Of wider significance are the associative algebras introduced in 1878 by Clifford [15] (pp. 266–276) as a common generalization of quaternions and Grassmann algebra. *Clifford algebras* were used by Lipschitz (1886) to represent orthogonal transformations in n-dimensional space. There is an extensive discussion of Clifford algebras in Deheuvels [20]. For their applications in physics, see Salingaros and Wene [64].

Proposition 32 has many uses. The proof given here is extracted from Nagahara and Tominaga [55].

It was proved by both Kervaire (1958) and Milnor (1958) that if a division algebra A (not necessarily associative) contains the real field $\mathbb{R}$ in its centre and is of finite dimension as a vector space over $\mathbb{R}$, then this dimension must be 1,2,4 or 8 (but the algebra need not be isomorphic to $\mathbb{R}$, $\mathbb{C}$, $\mathbb{H}$ or $\mathbb{O}$). All known proofs use deep results from algebraic topology, which was first applied to the problem by H. Hopf (1940). For more information about the proof, see Chapter 11 (by Hirzebruch) of Ebbinghaus *et al*. [24].

When is the product of two sums of squares again a sum of squares? To make the question precise, call a triple (r, s, t) of positive integers 'admissible' if there exist real numbers $\rho_{ijk} (1 \leq i \leq t, 1 \leq j \leq r, 1 \leq k \leq s)$ such that, for every $x = (\xi_1, \ldots, \xi_r) \in \mathbb{R}^r$ and every $y = (\eta_1, \ldots, \eta_s) \in \mathbb{R}^s$,

$$(\xi_1^2 + \cdots + \xi_r^2)(\eta_1^2 + \cdots + \eta_s^2) = \zeta_1^2 + \cdots + \zeta_t^2,$$

where

$$\zeta_i = \sum_{j=1}^{r} \sum_{k=1}^{s} \rho_{ijk} \xi_j \eta_k.$$

The question then becomes, which triples (r, s, t) are admissible? It is obvious that $(1, 1, 1)$ is admissible and the relation $n(x)n(y) = n(xy)$ for the norms of complex numbers, quaternions and octonions shows that (t, t, t) is admissible also for $t = 2, 4, 8$. It was proved by Hurwitz (1898) that (t, t, t) is admissible for no other values of t. A survey of the general problem is given by Shapiro [65].

General introductions to algebra are provided by Birkhoff and MacLane [7] and Herstein [35]. More extended treatments are given in Jacobson [41] and Lang [48].

The theory of groups is treated in M. Hall [29] and Rotman [60]. An especially significant class of groups is studied in Humphreys [38].

If H is a subgroup of a finite group G, then it is possible to choose a system of left coset representatives of H which is also a system of right coset representatives. This interesting, but not very useful, fact belongs to combinatorics rather than to group theory. We mention it because it was the motivation for the theorem of P. Hall (1935) on *systems of distinct representatives*, also known as the 'marriage theorem'. Further developments are described in Mirsky [53]. For quantitative versions, with applications to operations research, see Ford and Fulkerson [27].

The theory of rings separates into two parts. Noncommutative ring theory, which now incorporates the structure theory of associative algebras, is studied in the books of Herstein [36], Kasch [42] and Lam [46]. Commutative ring theory, which grew out of algebraic number theory and algebraic geometry, is studied in Atiyah and Macdonald [4] and Kunz [45].

Field theory was established as an independent subject of study in 1910 by Steinitz [68]. The books of Jacobson [41] and Lang [48] treat also the more recent theory of ordered fields, due to Artin and Schreier (1927).

Fields and groups are connected with one another by *Galois theory*. This subject has its origin in attempts to solve polynomial equations 'by radicals'. The founder of the subject is really Lagrange (1770/1). By developing his ideas, Ruffini (1799) and Abel (1826) showed that polynomial equations of degree greater than 4 cannot, in general, be solved by radicals. Abel (1829) later showed that polynomial equations *can* be solved by radicals if their 'Galois group' is commutative. In honour of this result, commutative groups are often called *abelian*.

Galois (1831, published posthumously in 1846) introduced the concept of normal subgroup and stated a necessary and sufficient condition for a polynomial equation to be solvable by radicals. The significance of Galois theory today lies not in this result, despite its historical importance, but in the much broader 'fundamental theorem of Galois theory'. In the form given it by Dedekind (1894) and Artin (1944), this establishes a correspondence between extension fields and groups of automorphisms, and provides a framework for the solution of a number of algebraic problems.

Morandi [54] and Rotman [61] give modern accounts of Galois theory. The historical development is traced in Kiernan [43]. In recent years attention has focussed on the problem of determining which finite groups occur as Galois groups over a given field; for an introductory account, see Matzat [51].

Some texts on linear algebra and matrix theory are Halmos [31], Horn and Johnson [37], Mal'cev [50] and Gantmacher [28].

The older literature on associative algebras is surveyed in Cartan [13]. The texts on noncommutative rings cited above give modern introductions.

A vast number of characterizations of inner product spaces, in addition to the parallelogram law, is given in Amir [3]. The theory of Hilbert space is treated in the books of Riesz and Sz.-Nagy [58] and Akhiezer and Glazman [2]. For its roots in the theory of integral equations, see Hellinger and Toeplitz [33]. Almost periodic functions are discussed from different points of view in Bohr [9], Corduneanu [17] and Maak [49]. The convergence of Fourier series is treated in Zygmund [73], for example.

12 Selected References

[1] L.V. Ahlfors, *Complex analysis*, 3rd ed., McGraw-Hill, New York, 1978.

[2] N.I. Akhiezer and I.M. Glazman, *Theory of linear operators in Hilbert space*, English transl. by E.R. Dawson based on 3rd Russian ed., Pitman, London, 1981.

[3] D. Amir, *Characterizations of inner product spaces*, Birkhäuser, Basel, 1986.

[4] M.F. Atiyah and I.G. Macdonald, *Introduction to commutative algebra*, Addison-Wesley, Reading, Mass., 1969.

[5] F.V. Atkinson, The reversibility of a differentiable mapping, *Canad. Math. Bull.* **4** (1961), 161–181.

[6] G. Birkhoff, *Lattice theory*, corrected reprint of 3rd ed., American Mathematical Society, Providence, R.I., 1979.

[7] G. Birkhoff and S. MacLane, *A survey of modern algebra*, 3rd ed., Macmillan, New York, 1965.

[8] F. van der Blij, History of the octaves, *Simon Stevin* **34** (1961), 106–125.

[9] H. Bohr, *Almost periodic functions*, English transl. by H. Cohn and F. Steinhardt, Chelsea, New York, 1947.
[10] G. Boole, *An investigation of the laws of thought, on which are founded the mathematical theories of logic and probability*, reprinted, Dover, New York, 1957. [Original edition, 1854]
[11] C. Caratheodory, *Theory of functions of a complex variable*, English transl. by F. Steinhardt, 2 vols., 2nd ed., Chelsea, New York, 1958/1960.
[12] G. Cardano, *The great art or the rules of algebra*, English transl. by T.R. Witmer, M.I.T. Press, Cambridge, Mass., 1968. [Latin original, 1545]
[13] E. Cartan, Nombres complexes, *Encyclopédie des sciences mathématiques, Tome* I, *Fasc*. 4, *Art*. I.5, Gauthier-Villars, Paris, 1908. [Reprinted in *Oeuvres complètes, Partie* II, *Vol*. 1, pp. 107–246.]
[14] G. Chichilnisky, Topology and invertible maps, *Adv. inAppl. Math.* **21** (1998), 113–123.
[15] W.K. Clifford, *Mathematical Papers*, reprinted, Chelsea, New York, 1968.
[16] E.A. Coddington and N. Levinson, *Theory of ordinary differential equations*, McGraw-Hill, New York, 1955.
[17] C. Corduneanu, *Almost periodic functions*, English transl. by G. Berstein and E. Tomer, Interscience, New York, 1968.
[18] T. Dantzig, *Number*: *The language of science*, 4th ed., Pi Press, Indianapolis, IN, 2005.
[19] R. Dedekind, *Essays on the theory of numbers*, English transl. by W.W. Beman, reprinted, Dover, New York, 1963.
[20] R. Deheuvels, *Formes quadratiques et groupes classiques*, Presses Universitaires de France, Paris, 1981.
[21] J. Dieudonné, *Foundations of modern analysis*, enlarged reprint, Academic Press, New York, 1969.
[22] J. Dieudonné *et al*., *Abrégé d'histoire des mathématiques 1700–1900*, reprinted, Hermann, Paris, 1996.
[23] J. Dugundji and A. Granas, *Fixed point theory* I, PWN, Warsaw, 1982.
[24] H.-D. Ebbinghaus *et al.*, *Numbers*, English transl. of 2nd German ed. by H.L.S. Orde, Springer-Verlag, New York, 1990.
[25] L. Euler, *Introduction to analysis of the infinite*, *Book* I, English transl. by J.D. Blanton, Springer-Verlag, New York, 1988.
[26] W. Fleming, *Functions of several variables*, 2nd ed., Springer-Verlag, New York, 1977.
[27] L.R. Ford Jr. and D.R. Fulkerson, *Flows in networks*, Princeton University Press, Princeton, N.J., 1962.
[28] F.R. Gantmacher, *The theory of matrices*, English transl. by K.A. Hirsch, 2 vols., Chelsea, New York, 1959.
[29] M. Hall, *The theory of groups*, reprinted, Chelsea, New York, 1976.
[30] P.R. Halmos, *Lectures on Boolean algebras*, Van Nostrand, Princeton, N.J., 1963.
[31] P.R. Halmos, *Finite-dimensional vector spaces*, 2nd ed., reprinted, Springer-Verlag, New York, 1974.
[32] G.G. Hamedani and G.G. Walter, A fixed point theorem and its application to the central limit theorem, *Arch. Math.* **43** (1984), 258–264.
[33] E. Hellinger and O. Toeplitz, *Integralgleichungen und Gleichungen mit unendlichvielen Unbekannten*, reprinted, Chelsea, New York, 1953. [Original edition, 1928]
[34] L. Henkin, On mathematical induction, *Amer. Math. Monthly* **67** (1960), 323–338.
[35] I.N. Herstein, *Topics in algebra*, reprinted, Wiley, London, 1976.
[36] I.N Herstein, *Noncommutative rings*, reprinted, Mathematical Association of America, Washington, D.C., 1994.
[37] R.A. Horn and C.A. Johnson, *Matrix analysis*, corrected reprint, Cambridge University Press, 1990.

[38] J.E. Humphreys, *Reflection groups and Coxeter groups*, Cambridge University Press, 1990.
[39] E.V. Huntingdon, Boolean algebra: A correction, *Trans. Amer. Math. Soc.* **35** (1933), 557–558.
[40] J. Jachymski, A short proof of the converse to the contraction principle and some related results, *Topol. Methods Nonlinear Anal.* **15** (2000), 179–186.
[41] N. Jacobson, *Basic Algebra I,II*, 2nd ed., Freeman, New York, 1985/1989.
[42] F. Kasch, *Modules and rings*, English transl. by D.A.R. Wallace, Academic Press, London, 1982.
[43] B.M. Kiernan, The development of Galois theory from Lagrange to Artin, *Arch. Hist. Exact Sci.* **8** (1971), 40–154.
[44] W. Kulpa, The Poincaré–Miranda theorem, *Amer. Math. Monthly* **104** (1997), 545–550.
[45] E. Kunz, *Introduction to commutative algebra and algebraic geometry*, English transl. by M. Ackerman, Birkhäuser, Boston, Mass., 1985.
[46] T.Y. Lam, *A first course in noncommutative rings*, Springer-Verlag, New York, 1991.
[47] E. Landau, *Foundations of analysis*, English transl. by F. Steinhardt, 3rd ed., Chelsea, New York, 1966. [German original, 1930]
[48] S. Lang, *Algebra*, corrected reprint of 3rd ed., Addison-Wesley, Reading, Mass., 1994.
[49] W. Maak, *Fastperiodische Funktionen*, Springer-Verlag, Berlin, 1950.
[50] A.I. Mal'cev, *Foundations of linear algebra*, English transl. by T.C. Brown, Freeman, San Francisco, 1963.
[51] B.H. Matzat, Über das Umkehrproblem der Galoisschen Theorie, *Jahresber. Deutsch. Math.-Verein.* **90** (1988), 155–183.
[52] K. Menninger, *Number words and number symbols*, English transl. by P. Broneer, M.I.T. Press, Cambridge, Mass., 1969.
[53] L. Mirsky, *Transversal theory*, Academic Press, London, 1971.
[54] P. Morandi, *Field and Galois theory*, Springer, New York, 1996.
[55] T. Nagahara and H. Tominaga, Elementary proofs of a theorem of Wedderburn and a theorem of Jacobson, *Abh. Math. Sem. Univ. Hamburg* **41** (1974), 72–74.
[56] R. Narasimhan, *Complex analysis in one variable*, Birkhäuser, Boston, Mass., 1985.
[57] A.H. Read, Higher derivatives of analytic functions from the standpoint of functional analysis, *J. London Math. Soc.* **36** (1961), 345–352.
[58] F. Riesz and B. Sz.-Nagy, *Functional analysis*, English transl. by L.F. Boron of 2nd French ed., F. Ungar, New York, 1955.
[59] H. Rothe, Systeme geometrischer Analyse, *Encyklopädie der Mathematischen Wissenschaften* III 1.2, pp. 1277–1423, Teubner, Leipzig, 1914–1931.
[60] J.J. Rotman, *An introduction to the theory of groups*, 4th ed., Springer-Verlag, New York, 1995.
[61] J. Rotman, *Galois theory*, 2nd ed., Springer-Verlag, New York, 1998.
[62] S. Rudeanu, *Boolean functions and equations*, North-Holland, Amsterdam, 1974.
[63] W. Rudin, *Principles of mathematical analysis*, 3rd ed., McGraw-Hill, New York, 1976.
[64] N.A. Salingaros and G.P. Wene, The Clifford algebra of differential forms, *Acta Appl. Math.* **4** (1985), 271–292.
[65] D.B. Shapiro, Products of sums of squares, *Exposition. Math.* **2** (1984), 235–261.
[66] R. Sikorski, *Boolean algebras*, 3rd ed., Springer-Verlag, New York, 1969.
[67] T.A. Springer and F.D. Veldkamp, *Octonions, Jordan algebras, and exceptional groups*, Springer, Berlin, 2000.
[68] E. Steinitz, *Algebraische Theorie der Körper*, reprinted, Chelsea, New York, 1950.
[69] M.H. Stone, The representation of Boolean algebras, *Bull. Amer. Math. Soc.* **44** (1938), 807–816.
[70] K.D. Stroyan and W.A.J. Luxemburg, *Introduction to the theory of infinitesimals*, Academic Press, New York, 1976.

[71] A. Sudbery, Quaternionic analysis, *Math. Proc. Cambridge Philos. Soc.* **85** (1979), 199–225.
[72] C.T.C. Wall, *A geometric introduction to topology*, reprinted, Dover, New York, 1993.
[73] A. Zygmund, *Trigonometric series*, 3rd ed., Cambridge University Press, 2003.

Additional References

J.C. Baez, The octonions, *Bull. Amer. Math. Soc.* (*N.S.*) **39** (2002), 145–205.
J.H. Conway and D.A. Smith, *On quaternions and octonions: their geometry, arithmetic and symmetry*, A.K. Peters, Natick, Mass., 2003.

II

Divisibility

1 Greatest Common Divisors

In the set $\mathbb{N}$ of all positive integers we can perform two basic operations: addition and multiplication. In this chapter we will be primarily concerned with the second operation.

Multiplication has the following properties:

(M1) *if* $ab = ac$, *then* $b = c$; (cancellation law)
(M2) $ab = ba$ *for all* a, b; (commutative law)
(M3) $(ab)c = a(bc)$ *for all* a, b, c; (associative law)
(M4) $1a = a$ *for all* a. (identity element)

For any $a, b \in \mathbb{N}$ we say that b *divides* a, or that b is a *factor* of a, or that a is a *multiple* of b if $a = ba'$ for some $a' \in \mathbb{N}$. We write $b|a$ if b divides a and $b \nmid a$ if b does not divide a. For example, $2|6$, since $6 = 2 \times 3$, but $4 \nmid 6$. (We sometimes use $\times$ instead of $\cdot$ for the product of positive integers.) The following properties of divisibility follow at once from the definition:

(i) $a|a$ *and* $1|a$ *for every* a;
(ii) *if* $b|a$ *and* $c|b$, *then* $c|a$;
(iii) *if* $b|a$, *then* $b|ac$ *for every* c;
(iv) $bc|ac$ *if and only if* $b|a$;
(v) *if* $b|a$ *and* $a|b$, *then* $b = a$.

For any $a, b \in \mathbb{N}$ we say that d is a *common divisor* of a and b if $d|a$ and $d|b$. We say that a common divisor d of a and b is a *greatest common divisor* if every common divisor of a and b divides d. The greatest common divisor of a and b is uniquely determined, if it exists, and will be denoted by (a, b).

The greatest common divisor of a and b is indeed the *numerically* greatest common divisor. However, it is preferable not to define greatest common divisors in this way, since the concept is then available for algebraic structures in which there is no relation of magnitude and only the operation of multiplication is defined.

W.A. Coppel, *Number Theory: An Introduction to Mathematics*, Universitext,
DOI: 10.1007/978-0-387-89486-7_2, © Springer Science + Business Media, LLC 2009

Proposition 1 *Any $a, b \in \mathbb{N}$ have a greatest common divisor (a, b).*

Proof Without loss of generality we may suppose $a \geq b$. If b divides a, then $(a, b) = b$. Assume that there exists a pair a, b without greatest common divisor and choose one for which a is a minimum. Then $1 < b < a$, since b does not divide a. Since also $1 \leq a - b < a$, the pair $a - b, b$ has a greatest common divisor d. Since any common divisor of a and b divides $a - b$, and since d divides $(a - b) + b = a$, it follows that d is a greatest common divisor of a and b. But this is a contradiction. □

The proof of Proposition 1 uses not only the multiplicative structure of the set $\mathbb{N}$, but also its ordering and additive structure. To see that there is a reason for this, consider the set S of all positive integers of the form $4k + 1$. The set S is closed under multiplication, since

$$(4j + 1)(4k + 1) = 4(4jk + j + k) + 1,$$

and we can define divisibility and greatest common divisors in S by simply replacing $\mathbb{N}$ by S in our previous definitions. However, although the elements 693 and 189 of S have the common divisors 9 and 21, they have no greatest common divisor according to this definition.

In the following discussion we use the result of Proposition 1, but make no further appeal to either addition or order.

For any $a, b \in \mathbb{N}$ we say that h is a *common multiple* of a and b if $a|h$ and $b|h$. We say that a common multiple h of a and b is a *least common multiple* if h divides every common multiple of a and b. The least common multiple of a and b is uniquely determined, if it exists, and will be denoted by $[a, b]$.

It is evident that, for every a,

$$(a, 1) = 1, \quad [a, 1] = a,$$
$$(a, a) = a = [a, a].$$

Proposition 2 *Any $a, b \in \mathbb{N}$ have a least common multiple $[a, b]$. Moreover,*

$$(a, b)[a, b] = ab.$$

Furthermore, for all $a, b, c \in \mathbb{N}$,

$$(ac, bc) = (a, b)c, \qquad [ac, bc] = [a, b]c,$$
$$([a, b], [a, c]) = [a, (b, c)], \quad [(a, b), (a, c)] = (a, [b, c]).$$

Proof We show first that $(ac, bc) = (a, b)c$. Put $d = (a, b)$. Clearly cd is a common divisor of ac and bc, and so $(ac, bc) = qcd$ for some $q \in \mathbb{N}$. Thus $ac = qcda'$, $bc = qcdb'$ for some $a', b' \in \mathbb{N}$. It follows that $a = qda'$, $b = qdb'$. Thus qd is a common divisor of a and b. Hence qd divides d, which implies $q = 1$.

If g is any common multiple of a and b, then ab divides ga and gb, and hence ab also divides (ga, gb). But, by what we have just proved,

$$(ga, gb) = (a, b)g = dg.$$

Hence $h := ab/d$ divides g. Since h is clearly a common multiple of a and b, it follows that $h = [a, b]$. Replacing a, b by ac, bc, we now obtain

$$[ac, bc] = acbc/(ac, bc) = abc/(a, b) = hc.$$

If we put

$$A = ([a, b], [a, c]), \quad B = [a, (b, c)],$$

then by what we have already proved,

$$A = (ab/(a, b), ac/(a, c)),$$
$$B = a(b, c)/(a, (b, c)) = (ab/(a, (b, c)), ac/(a, (b, c))).$$

Since any common divisor of $ab/(a, b)$ and $ac/(a, c)$ is also a common divisor of $ab/(a, (b, c))$ and $ac/(a, (b, c))$, it follows that A divides B. On the other hand, a divides A, since a divides $[a, b]$ and $[a, c]$, and similarly (b, c) divides A. Hence B divides A. Thus $B = A$.

The remaining statement of the proposition is proved in the same way, with greatest common divisors and least common multiples interchanged. □

The last two statements of Proposition 2 are referred to as the distributive laws, since if the greatest common divisor and least common multiple of a and b are denoted by $a \wedge b$ and $a \vee b$ respectively, they take the form

$$(a \vee b) \wedge (a \vee c) = a \vee (b \wedge c), \quad (a \wedge b) \vee (a \wedge c) = a \wedge (b \vee c).$$

Properties (i), (ii) and (v) at the beginning of the section say that divisibility is a *partial ordering* of the set $\mathbb{N}$ with 1 as least element. The existence of greatest common divisors and least common multiples says that $\mathbb{N}$ is a *lattice* with respect to this partial ordering. The distributive laws say that $\mathbb{N}$ is actually a *distributive* lattice.

We say that $a, b \in \mathbb{N}$ are *relatively prime*, or *coprime*, if $(a, b) = 1$. Divisibility properties in this case are much simpler:

Proposition 3 *For any $a, b, c \in \mathbb{N}$ with $(a, b) = 1$,*

(i) *if $a|c$ and $b|c$, then $ab|c$;*
(ii) *if $a|bc$, then $a|c$;*
(iii) $(a, bc) = (a, c)$*;*
(iv) *if also $(a, c) = 1$, then $(a, bc) = 1$;*
(v) $(a^m, b^n) = 1$ *for all $m, n \geq 1$.*

Proof To prove (i), note that $[a, b]$ divides c and $[a, b] = ab$. To prove (ii), note that a divides $(ac, bc) = (a, b)c = c$. To prove (iii), note that any common divisor of a and bc divides c, by (ii). Obviously (iii) implies (iv), and (v) follows by induction. □

Proposition 4 *If $a, b \in \mathbb{N}$ and $(a, b) = 1$, then any divisor of ab can be uniquely expressed in the form de, where $d|a$ and $e|b$. Conversely, any product of this form is a divisor of ab.*

Proof The proof is based on Proposition 3. Suppose c divides ab and put $d = (a, c)$, $e = (b, c)$. Then $(d, e) = 1$ and hence de divides c. If $a = da'$ and $c = dc'$, then $(a', c') = 1$ and $e|c'$. On the other hand, $c'|a'b$ and hence $c'|b$. Since $e = (b, c)$, it follows that $c' = e$ and $c = de$.

Suppose $de = d'e'$, where d, d' divide a and e, e' divide b. Then $d|d'$, since $(d, e') = 1$, and similarly $d'|d$, since $(d', e) = 1$. Hence $d' = d$ and $e' = e$.

The final statement of the proposition is obvious. □

It follows from Proposition 4 that if $c^n = ab$, where $(a, b) = 1$, then $a = d^n$ and $b = e^n$ for some $d, e \in \mathbb{N}$.

The greatest common divisor and least common multiple of any finite set of elements of $\mathbb{N}$ may be defined in the same way as for sets of two elements. By induction we easily obtain:

Proposition 5 *Any $a_1, \ldots, a_n \in \mathbb{N}$ have a greatest common divisor $(a_1, \ldots, a_n)$ and a least common multiple $[a_1, \ldots, a_n]$. Moreover,*

(i) $(a_1, a_2, \ldots, a_n) = (a_1, (a_2, \ldots, a_n))$, $[a_1, a_2, \ldots, a_n] = [a_1, [a_2, \ldots, a_n]]$;
(ii) $(a_1 c, \ldots, a_n c) = (a_1, \ldots, a_n)c$, $[a_1 c, \ldots, a_n c] = [a_1, \ldots, a_n]c$;
(iii) $(a_1, \ldots, a_n) = a/[a/a_1, \ldots, a/a_n]$, $[a_1, \ldots, a_n] = a/(a/a_1, \ldots, a/a_n)$, *where* $a = a_1 \cdots a_n$.

We can use the distributive laws to show that

$$([a, b], [a, c], [b, c]) = [(a, b), (a, c), (b, c)].$$

In fact the left side is equal to $\{a \vee (b \wedge c)\} \wedge (b \vee c)$, whereas the right side is equal to

$$\begin{aligned}(b \wedge c) \vee \{a \wedge (b \vee c)\} &= \{(b \wedge c) \vee a\} \wedge \{(b \wedge c) \vee (b \vee c)\} \\ &= \{a \vee (b \wedge c)\} \wedge (b \vee c).\end{aligned}$$

If

$$a = (a_1, \ldots, a_m), \quad b = (b_1, \ldots, b_n),$$

then ab is the greatest common divisor of all products $a_j b_k$, since $(a_j b_1, \ldots, a_j b_n) = a_j b$ and $(a_1 b, \ldots, a_m b) = ab$.

Similarly, if

$$a = [a_1, \ldots, a_m], \quad b = [b_1, \ldots, b_n],$$

then ab is the least common multiple of all products $a_j b_k$.

It is easily shown by induction that if $(a_i, a_j) = 1$ for $1 \leq i < j \leq m$, then

$$(a_1 \cdots a_m, c) = (a_1, c) \cdots (a_m, c), [a_1 \cdots a_m, c] = [a_1, \ldots, a_m, c].$$

Proposition 6 *If $a \in \mathbb{N}$ has two factorizations*

$$a = b_1 \cdots b_m = c_1 \cdots c_n,$$

then these factorizations have a common refinement, i.e. there exist $d_{jk} \in \mathbb{N}$ $(1 \leq j \leq m, 1 \leq k \leq n)$ such that

$$b_j = \prod_{k=1}^{n} d_{jk}, \quad c_k = \prod_{j=1}^{m} d_{jk}.$$

Proof We show first that if $a = a_1 \cdots a_n$ and $d|a$, then $d = d_1 \cdots d_n$, where $d_i|a_i$ $(1 \leq i \leq n)$. We may suppose that $n > 1$ and that the assertion holds for products of less than n elements of $\mathbb{N}$. Put $a' = a_1 \cdots a_{n-1}$ and $d' = (a', d)$. Then $d' = d_1 \cdots d_{n-1}$, where $d_i|a_i (1 \leq i < n)$. Moreover $a'' = a'/d'$ and $d'' = d/d'$ are coprime. Since $d'' = d/d'$ divides $a''a_n = a/d'$, the greatest common divisor $a_n = (a_n a'', a_n d'')$ is divisible by d''. Thus we can take $d_n = d''$.

We return now to the proposition. Since $c_1 | \prod_j b_j$, we can write $c_1 = \prod_j d_{j1}$, where $d_{j1}|b_j$. Put $b'_j = b_j/d_{j1}$. Then

$$\prod_j b'_j = a/c_1 = c_2 \cdots c_n.$$

Hence we can write $c_2 = \prod_j d_{j2}$, where $d_{j2}|b'_j$. Proceeding in this way, we obtain factorizations $c_k = \prod_j d_{jk}$ such that $\prod_k d_{jk}$ divides b_j. In fact, since

$$\prod_{j,k} d_{jk} = a = \prod_j b_j,$$

we must have $b_j = \prod_k d_{jk}$. □

Instead of defining divisibility and greatest common divisors in the set $\mathbb{N}$ of all positive integers, we can define them in the set $\mathbb{Z}$ of all integers by simply replacing $\mathbb{N}$ by $\mathbb{Z}$ in the previous definitions. The properties (i)–(v) at the beginning of this section continue to hold, provided that in (iv) we require $c \neq 0$ and in (v) we alter the conclusion to $b = \pm a$. We now list some additional properties:

(i)′ $a|0$ *for every* a;
(ii)′ *if* $0|a$, *then* $a = 0$;
(iii)′ *if* $c|a$ *and* $c|b$, *then* $c|ax + by$ *for all* x, y.

Greatest common divisors and least common multiples still exist, but uniqueness holds only up to sign. With this understanding, Propositions 2–4 continue to hold, and so also do Propositions 5 and 6 if we require $a \neq 0$. It is evident that, for every a,

$$(a, 0) = a, \quad [a, 0] = 0.$$

More generally, we can define divisibility in any *integral domain*, i.e. a commutative ring in which $a \neq 0$ and $b \neq 0$ together imply $ab \neq 0$. The properties (i)–(v) at the beginning of the section continue to hold, provided that in (iv) we require $c \neq 0$ and in (v) we alter the conclusion to $b = ua$, where u is a *unit*, i.e. $u|1$. The properties (i)′–(iii)′ above also remain valid.

We define a *GCD domain* to be an integral domain in which any pair of elements has a greatest common divisor. This implies that any pair of elements also has a least common multiple. Uniqueness now holds only up to unit multiples. With this understanding Propositions 2–6 continue to hold in any GCD domain in the same way as for $\mathbb{Z}$.

An important example, which we will consider in Section 3, of a GCD domain other than $\mathbb{Z}$ is the *polynomial ring* $K[t]$, consisting of all polynomials in t with coefficients from an arbitrary field K. The units in this case are the nonzero elements of K.

Another example, which we will meet in §4 of Chapter VI, is the valuation ring R of a non-archimedean valued field. In this case, for any $a, b \in R$, either $a|b$ or $b|a$ and so (a, b) is either a or b.

In the same way that the ring $\mathbb{Z}$ of integers may be embedded in the field $\mathbb{Q}$ of rational numbers, any integral domain R may be embedded in a field K, its *field of fractions*, so that any nonzero $c \in K$ has the form $c = ab^{-1}$, where $a, b \in R$ and $b \neq 0$. If R is a GCD domain we can further require $(a, b) = 1$, and a, b are then uniquely determined apart from a common unit multiple. The field of fractions of the polynomial ring $K[t]$ is the field $K(t)$ of *rational functions*.

In our discussion of divisibility so far we have avoided all mention of prime numbers. A positive integer $a \neq 1$ is said to be *prime* if 1 and a are its only positive divisors, and otherwise is said to be *composite*.

For example, 2, 3 and 5 are primes, but $4 = 2 \times 2$ and $6 = 2 \times 3$ are composite. The significance of the primes is that, as far as multiplication is concerned, they are the 'atoms' and the composite integers are the 'molecules'. This is made precise in the following so-called *fundamental theorem of arithmetic*:

Proposition 7 *If $a \in \mathbb{N}$ and $a \neq 1$, then a can be represented as a product of finitely many primes. Moreover, the representation is unique, except for the order of the factors.*

Proof Assume, on the contrary, that some composite $a_1 \in \mathbb{N}$ is not a product of finitely many primes. Since a_1 is composite, it has a factorization $a_1 = a_2 b_2$, where $a_2, b_2 \in \mathbb{N}$ and $a_2, b_2 \neq 1$. At least one of a_2, b_2 must be composite and not a product of finitely many primes, and we may choose the notation so that a_2 has these properties. The preceding argument can now be repeated with a_2 in place of a_1. Proceeding in this way, we obtain an infinite sequence (a_k) of positive integers such that a_{k+1} divides a_k and $a_{k+1} \neq a_k$ for each $k \geq 1$. But then the sequence (a_k) has no least element, which contradicts Proposition I.3.

Suppose now that

$$a = p_1 \cdots p_m = q_1 \cdots q_n$$

are two representations of a as a product of primes. Then, by Proposition 6, there exist $d_{jk} \in \mathbb{N}$ $(1 \leq j \leq m, 1 \leq k \leq n)$ such that

$$p_j = \prod_{k=1}^{n} d_{jk}, \quad q_k = \prod_{j=1}^{m} d_{jk}.$$

Since p_1 is a prime, we must have $d_{1k_1} = p_1$ for some $k_1 \in \{1, \ldots, n\}$, and since q_{k_1} is a prime, we must have $q_{k_1} = d_{1k_1} = p_1$. The same argument can now be applied to

$$a' = \prod_{j \neq 1} p_j = \prod_{k \neq k_1} q_k.$$

It follows that $m = n$ and $q_1, \ldots, q_n$ is a permutation of $p_1, \ldots, p_m$. □

It should be noted that factorization into primes would not be unique if we admitted 1 as a prime. The fundamental theorem of arithmetic may be reformulated in the following way: any $a \in \mathbb{N}$ can be uniquely represented in the form

$$a = \prod_p p^{\alpha_p},$$

where p runs through the primes and the α_p are non-negative integers, only finitely many of which are nonzero. It is easily seen that if $b \in \mathbb{N}$ has the analogous representation

$$b = \prod_p p^{\beta_p},$$

then $b|a$ if and only if $\beta_p \leq \alpha_p$ for all p. It follows that the greatest common divisor and least common multiple of a and b have the representations

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

where

$$\gamma_p = \min\{\alpha_p, \beta_p\}, \quad \delta_p = \max\{\alpha_p, \beta_p\}.$$

The fundamental theorem of arithmetic extends at once from $\mathbb{N}$ to $\mathbb{Q}$: any nonzero rational number a can be uniquely represented in the form

$$a = u \prod_p p^{\alpha_p},$$

where $u = \pm 1$ is a unit, p runs through the primes and the α_p are integers (not necessarily non-negative), only finitely many of which are nonzero.

The following property of primes was already established in Euclid's *Elements* (Book VII, Proposition 30):

Proposition 8 *If p is a prime and $p|bc$, then $p|b$ or $p|c$.*

Proof If p does not divide b, we must have $(p, b) = 1$. But then p divides c, by Proposition 3(ii). □

The property in Proposition 8 actually characterizes primes. For if a is composite, then $a = bc$, where $b, c \neq 1$. Thus $a|bc$, but $a \nmid b$ and $a \nmid c$.

We consider finally the extension of these notions to an arbitrary integral domain R. For any nonzero $a, b \in R$, we say that a divisor b of a is a *proper divisor* if a does not divide b (i.e., if a and b do not differ only by a unit factor). We say that $p \in R$ is *irreducible* if p is neither zero nor a unit and if every proper divisor of p is a unit. We say that $p \in R$ is *prime* if p is neither zero nor a unit and if $p|bc$ implies $p|b$ or $p|c$.

By what we have just said, the notions of 'prime' and 'irreducible' coincide if $R = \mathbb{Z}$, and the same argument applies if R is any GCD domain. However, in an arbitrary integral domain R, although any prime element is irreducible, an irreducible element need not be prime. (For example, in the integral domain R consisting of all complex numbers of the form $a + b\sqrt{-5}$, where $a, b \in \mathbb{Z}$, it may be seen that

$6 = 2 \times 3 = (1 + \sqrt{-5})(1 - \sqrt{-5})$ has two essentially distinct factorizations into irreducibles, and thus none of these irreducibles is prime.)

The proof of Proposition 7 shows that, in an arbitrary integral domain R, every element which is neither zero nor a unit can be represented as a product of finitely many irreducible elements if and only if the following *chain condition* is satisfied:

(#) *there exists no infinite sequence* (a_n) *of elements of* R *such that* a_{n+1} *is a proper divisor of* a_n *for every* n.

Furthermore, the representation is *essentially unique* (i.e. unique except for the order of the factors and for multiplying them by units) if and only if R is also a GCD domain.

An integral domain R is said to be *factorial* (or a 'unique factorization domain') if the 'fundamental theorem of arithmetic' holds in R, i.e. if every element which is neither zero nor a unit has such an essentially unique representation as a product of finitely many irreducibles. By the above remarks, an integral domain R is factorial if and only if it is a GCD domain satisfying the chain condition (#).

For future use, we define an element of a factorial domain to be *square-free* if it is neither zero nor a unit and if, in its representation as a product of irreducibles, no factor is repeated. In particular, a positive integer is square-free if and only if it is a nonempty product of distinct primes.

2 The Bézout Identity

If a, b are arbitrary integers with $a \neq 0$, then there exist unique integers q, r such that

$$b = qa + r, \quad 0 \leq r < |a|.$$

In fact qa is the greatest multiple of a which does not exceed b. The integers q and r are called the *quotient* and *remainder* in the 'division' of b by a.

(For $a > 0$ this was proved in Proposition I.14. It follows that if a and n are positive integers, any positive integer b less than a^n has a unique representation 'to the base a':

$$b = b_0 + b_1 a + \cdots + b_{n-1} a^{n-1},$$

where $0 \leq b_j < a$ for all j. In fact b_{n-1} is the quotient in the division of b by a^{n-1}, b_{n-2} is the quotient in the division of the remainder by a^{n-2}, and so on.)

If a, b are arbitrary integers with $a \neq 0$, then there exist also integers q, r such that

$$b = qa + r, \quad |r| \leq |a|/2.$$

In fact qa is the nearest multiple of a to b. Thus q and r are not uniquely determined if b is midway between two consecutive multiples of a.

Both these *division algorithms* have their uses. We will be impartial and merely use the fact that

$$b = qa + r, \quad |r| < |a|.$$

An *ideal* in the commutative ring $\mathbb{Z}$ of all integers is defined to be a nonempty subset J such that if $a, b \in J$ and $x, y \in \mathbb{Z}$, then also $ax + by \in J$.

For example, if $a_1, \ldots, a_n$ are given elements of $\mathbb{Z}$, then the set of all linear combinations $a_1x_1 + \cdots + a_nx_n$ with $x_1, \ldots, x_n \in \mathbb{Z}$ is an ideal, the ideal *generated* by $a_1, \ldots, a_n$. An ideal generated by a single element, i.e. the set of all multiples of that element, is said to be a *principal ideal*.

Lemma 9 *Any ideal J in the ring $\mathbb{Z}$ is a principal ideal.*

Proof If 0 is the only element of J, then 0 generates J. Otherwise there is a nonzero $a \in J$ with minimum absolute value. For any $b \in J$, we can write $b = qa + r$, for some $q, r \in \mathbb{Z}$ with $|r| < |a|$. By the definition of an ideal, $r \in J$ and so, by the definition of a, $r = 0$. Thus a generates J. □

Proposition 10 *Any $a, b \in \mathbb{Z}$ have a greatest common divisor $d = (a, b)$. Moreover, for any $c \in \mathbb{Z}$, there exist $x, y \in \mathbb{Z}$ such that*

$$ax + by = c$$

if and only if d divides c.

Proof Let J be the ideal generated by a and b. By Lemma 9, J is generated by a single element d. Since $a, b \in J$, d is a common divisor of a and b. On the other hand, since $d \in J$, there exist $u, v \in \mathbb{Z}$ such that $d = au + bv$. Hence any common divisor of a and b also divides d. Thus $d = (a, b)$. The final statement of the proposition follows immediately since, by definition, $c \in J$ if and only if there exist $x, y \in \mathbb{Z}$ such that $ax + by = c$. □

It is readily shown that if the 'linear Diophantine' equation $ax + by = c$ has a solution $x_0, y_0 \in \mathbb{Z}$, then all solutions $x, y \in \mathbb{Z}$ are given by the formula

$$x = x_0 + kb/d, \quad y = y_0 - ka/d,$$

where $d = (a, b)$ and k is an arbitrary integer.

Proposition 10 provides a new proof for the existence of greatest common divisors and, in addition, it shows that the greatest common divisor of two integers can be represented as a linear combination of them. This representation is usually referred to as the *Bézout identity*, although it was already known to Bachet (1624) and even earlier to the Hindu mathematicians Aryabhata (499) and Brahmagupta (628).

In exactly the same way that we proved Proposition 10 – or, alternatively, by induction from Proposition 10 – we can prove

Proposition 11 *Any finite set $a_1, \ldots, a_n$ of elements of $\mathbb{Z}$ has a greatest common divisor $d = (a_1, \ldots, a_n)$. Moreover, for any $c \in \mathbb{Z}$, there exist $x_1, \ldots, x_n \in \mathbb{Z}$ such that*

$$a_1x_1 + \cdots + a_nx_n = c$$

if and only if d divides c.

The proof which we gave for Proposition 10 is a pure existence proof – it does not help us to find the greatest common divisor. The following constructive proof was already given in Euclid's *Elements* (Book VII, Proposition 2). Let a, b be arbitrary

integers. Since $(0, b) = b$, we may assume $a \neq 0$. Then there exist integers q, r such that

$$b = qa + r, \quad |r| < |a|.$$

Put $a_0 = b$, $a_1 = a$ and repeatedly apply this procedure:

$$\begin{aligned} a_0 &= q_1 a_1 + a_2, \quad |a_2| < |a_1|, \\ a_1 &= q_2 a_2 + a_3, \quad |a_3| < |a_2|, \\ &\cdots \\ a_{N-2} &= q_{N-1} a_{N-1} + a_N, \quad |a_N| < |a_{N-1}|, \\ a_{N-1} &= q_N a_N. \end{aligned}$$

The process must eventually terminate as shown, because otherwise we would obtain an infinite sequence of positive integers with no least element. We claim that a_N is a greatest common divisor of a and b. In fact, working forwards from the first equation we see that any common divisor c of a and b divides each a_k and so, in particular, a_N. On the other hand, working backwards from the last equation we see that a_N divides each a_k and so, in particular, a and b.

The Bézout identity can also be obtained in this way, although Euclid himself lacked the necessary algebraic notation. Define sequences (x_k), (y_k) by the recurrence relations

$$x_{k+1} = x_{k-1} - q_k x_k, \quad y_{k+1} = y_{k-1} - q_k y_k \quad (1 \leq k < N),$$

with the starting values

$$x_0 = 0, \ \ x_1 = 1, \quad \text{resp. } y_0 = 1, \ y_1 = 0.$$

It is easily shown by induction that $a_k = ax_k + by_k$ and so, in particular, $a_N = ax_N + by_N$.

The Euclidean algorithm is quite practical. For example, the reader may use it to verify that 13 is the greatest common divisor of 2171 and 5317, and that

$$49 \times 5317 - 120 \times 2171 = 13.$$

However, the first proof given for Proposition 10 also has its uses: there is some advantage in separating the conceptual from the computational and the proof actually rests on more general principles, since there are quadratic number fields whose ring of integers is a 'principal ideal domain' that does not possess any Euclidean algorithm.

It is not visibly obvious that the binomial coefficients

$${}^{m+n}C_n = (m+1)\cdots(m+n)/1 \cdot 2 \cdot \cdots \cdot n$$

are integers for all positive integers m, n, although it is apparent from their combinatorial interpretation. However, the property is readily proved by induction, using the relation

$${}^{m+n}C_n = {}^{m+n-1}C_n + {}^{m+n-1}C_{n-1}.$$

Binomial coefficients have other arithmetic properties. Hermite observed that ${}^{m+n}C_n$ is divisible by the integers $(m+n)/(m,n)$ and $(m+1)/(m+1,n)$. In particular, the *Catalan numbers* $(n+1)^{-1}\,{}^{2n}C_n$ are integers. The following proposition is a substantial generalization of these results and illustrates the application of Proposition 10.

Proposition 12 *Let (a_n) be a sequence of nonzero integers such that, for all $m, n \geq 1$, every common divisor of a_m and a_n divides a_{m+n}, and every common divisor of a_m and a_{m+n} divides a_n. Then, for all $m, n \geq 1$,*

(i) $(a_m, a_n) = a_{(m,n)}$;
(ii) $A_{m,n} := a_{m+1} \cdots a_{m+n}/a_1 \cdots a_n \in \mathbb{Z}$;
(iii) *$A_{m,n}$ is divisible by $a_{m+n}/(a_m, a_n)$, by $a_{m+1}/(a_{m+1}, a_n)$ and by $a_{n+1}/(a_m, a_{n+1})$;*
(iv) $(A_{m,n-1}, A_{m+1,n}, A_{m-1,n+1}) = (A_{m-1,n}, A_{m+1,n-1}, A_{m,n+1})$.

Proof The hypotheses imply that

$$(a_m, a_n) = (a_m, a_{m+n}) \quad \text{for all } m, n \geq 1.$$

Since $a_m = (a_m, a_m)$, it follows by induction that $a_m | a_{km}$ for all $k \geq 1$. Moreover,

$$(a_{km}, a_{(k+1)m}) = a_m,$$

since every common divisor of a_{km} and $a_{(k+1)m}$ divides a_m.

Put $d = (m, n)$. Then $m = dm'$, $n = dn'$, where $(m', n') = 1$. Thus there exist integers u, v such that $m'u - n'v = 1$. By replacing u, v by $u + tn', v + tm'$ with any $t > \max\{|u|, |v|\}$, we may assume that u and v are both positive. Then

$$(a_{mu}, a_{nv}) = (a_{(n'v+1)d}, a_{n'vd}) = a_d.$$

Since a_d divides (a_m, a_n) and (a_m, a_n) divides (a_{mu}, a_{nv}), this implies $(a_m, a_n) = a_d$. This proves (i).

Since $a_1 | a_{m+1}$, it is evident that $A_{m,1} \in \mathbb{Z}$ for all $m \geq 1$. We assume that $n > 1$ and $A_{m,n} \in \mathbb{Z}$ for all smaller values of n and all $m \geq 1$. Since it is trivial that $A_{0,n} \in \mathbb{Z}$, we assume also that $m \geq 1$ and $A_{m,n} \in \mathbb{Z}$ for all smaller values of m. By Proposition 10, there exist $x, y \in \mathbb{Z}$ such that

$$a_m x + a_n y = a_{m+n},$$

since (a_m, a_n) divides a_{m+n}. Since

$$A_{m,n} = \frac{a_{m+1} \cdots a_{m+n}}{a_1 \cdots a_n} = \frac{a_m a_{m+1} \cdots a_{m+n-1}}{a_1 \cdots a_n} x + \frac{a_{m+1} \cdots a_{m+n-1}}{a_1 \cdots a_{n-1}} y,$$

our induction hypotheses imply that $A_{m,n} \in \mathbb{Z}$. This proves (ii).

Since

$$a_{m+n} A_{m,n-1} = a_n A_{m,n},$$

a_{m+n} divides $(a_n, a_{m+n}) A_{m,n}$ and, since $(a_n, a_{m+n}) = (a_m, a_n)$, this in turn implies that $a_{m+n}/(a_m, a_n)$ divides $A_{m,n}$.

Similarly, since

$$a_{m+1}A_{m+1,n} = a_{m+n+1}A_{m,n}, \quad a_{m+1}A_{m+1,n-1} = a_nA_{m,n},$$

a_{m+1} divides $(a_n, a_{m+n+1})A_{m,n}$ and, since $(a_n, a_{m+n+1}) = (a_{m+1}, a_n)$, it follows that $a_{m+1}/(a_{m+1}, a_n)$ divides $A_{m,n}$. In the same way, since

$$a_{n+1}A_{m,n+1} = a_{m+n+1}A_{m,n}, \quad a_{n+1}A_{m-1,n+1} = a_mA_{m,n},$$

a_{n+1} divides $(a_m, a_{m+n+1})A_{m,n}$ and hence $a_{n+1}/(a_m, a_{n+1})$ divides $A_{m,n}$. This proves (iii).

By multiplying by $a_1 \cdots a_{n+1}/a_{m+2} \cdots a_{m+n-1}$, we see that (iv) is equivalent to

$$\begin{aligned}&(a_na_{n+1}a_{m+1}, a_{n+1}a_{m+n}a_{m+n+1}, a_ma_{m+1}a_{m+n})\\ &\quad = (a_{n+1}a_ma_{m+1}, a_na_{n+1}a_{m+n}, a_{m+1}a_{m+n}a_{m+n+1}).\end{aligned}$$

Since here the two sides are interchanged when m and n are interchanged, it is sufficient to show that any common divisor e of the three terms on the right is also a common divisor of the three terms on the left. We have

$$\begin{aligned}(a_{n+1}a_ma_{m+1}, a_na_{n+1}a_{m+1}) = a_{n+1}a_{m+1}(a_m, a_n) &= a_{n+1}a_{m+1}(a_m, a_{m+n})\\ &= (a_{n+1}a_ma_{m+1}, a_{m+1}a_{n+1}a_{m+n}),\end{aligned}$$

and similarly

$$\begin{aligned}(a_na_{n+1}a_{m+n}, a_{n+1}a_{m+n}a_{m+n+1}) &= (a_na_{n+1}a_{m+n}, a_{m+1}a_{n+1}a_{m+n}),\\ (a_{m+1}a_{m+n}a_{m+n+1}, a_ma_{m+1}a_{m+n}) &= (a_{m+1}a_{m+n}a_{m+n+1}, a_{m+1}a_{n+1}a_{m+n}).\end{aligned}$$

Hence if we put $g = a_{m+1}a_{n+1}a_{m+n}$, then

$$(e, g) = (e, a_na_{n+1}a_{m+1}) = (e, a_{n+1}a_{m+n}a_{m+n+1}) = (e, a_ma_{m+1}a_{m+n})$$

and if we put $f = (e,g)$, then

$$1 = (e/f, a_na_{n+1}a_{m+1}/f) = (e/f, a_{n+1}a_{m+n}a_{m+n+1}/f) = (e/f, a_ma_{m+1}a_{m+n}/f).$$

Hence $(e/f, P/f^3) = 1$, where

$$P = a_na_{n+1}a_{m+1} \cdot a_{n+1}a_{m+n}a_{m+n+1} \cdot a_ma_{m+1}a_{m+n}.$$

But P is divisible by e^3, since we can also write

$$P = a_{n+1}a_ma_{m+1} \cdot a_na_{n+1}a_{m+n} \cdot a_{m+1}a_{m+n}a_{m+n+1}.$$

Hence the previous relation implies $e/f = 1$. Thus $e = f$ is a common divisor of $a_na_{n+1}a_{m+1}$, $a_{n+1}a_{m+n}a_{m+n+1}$ and $a_ma_{m+1}a_{m+n}$, as we wished to show. □

For the binomial coefficient case, i.e. $a_n = n$, the property (iv) of Proposition 12 was discovered empirically by Gould (1972) and then proved by Hillman and Hoggatt (1972). It states that if in the *Pascal triangle* one picks out the hexagon surrounding a particular element, then the greatest common divisor of three alternately

chosen vertices is equal to the greatest common divisor of the remaining three vertices. Hillman and Hoggatt also gave generalizations along the lines of Proposition 12.

The hypotheses of Proposition 12 are also satisfied if $a_n = q^n - 1$, for some integer $q > 1$, since in this case $a_{m+n} = a_m a_n + a_m + a_n$. The corresponding *q-binomial coefficients* were studied by Gauss and, as mentioned in Chapter XIII, they play a role in the theory of partitions.

We may also take (a_n) to be the sequence defined recurrently by

$$a_1 = 1, \quad a_2 = c, \quad a_{n+2} = ca_{n+1} + ba_n (n \geq 1),$$

where b and c are coprime positive integers. Indeed it is easily shown by induction that

$$(a_n, a_{n+1}) = (b, a_{n+1}) = 1 \quad \text{for all } n \geq 1.$$

By induction on m one may also show that

$$a_{m+n} = a_{m+1}a_n + ba_m a_{n-1} \quad \text{for all } m \geq 1,\ n > 1.$$

It follows that the hypotheses of Proposition 12 are satisfied. In particular, for $b = c = 1$, they are satisfied by the sequence of *Fibonacci numbers*.

We consider finally extensions of our results to more general algebraic structures. An integral domain R is said to be a *Bézout domain* if any $a, b \in R$ have a common divisor of the form $au + bv$ for some $u, v \in R$. Since such a common divisor is necessarily a greatest common divisor, any Bézout domain is a GCD domain. It is easily seen, by induction on the number of generators, that an integral domain is a Bézout domain if and only if every finitely generated ideal is a principal ideal. Thus Propositions 10 and 11 continue to hold if $\mathbb{Z}$ is replaced by any Bézout domain.

An integral domain R is said to be a *principal ideal domain* if every ideal is a principal ideal.

Lemma 13 *An integral domain R is a principal ideal domain if and only if it is a Bézout domain satisfying the chain condition*

(#) there exists no infinite sequence (a_n) of elements of R such that a_{n+1} is a proper divisor of a_n for every n.

Proof It is obvious that any principal ideal domain is a Bézout domain. Suppose R is a Bézout domain, but not a principal ideal domain. Then R contains an ideal J which is not finitely generated. Hence there exists a sequence (b_n) of elements of J such that b_{n+1} is not in the ideal J_n generated by $b_1, \ldots, b_n$. But J_n is a principal ideal. If a_n generates J_n, then a_{n+1} is a proper divisor of a_n for every n. Thus the chain condition is violated.

Suppose now that R is a Bézout domain containing a sequence (a_n) such that a_{n+1} is a proper divisor of a_n for every n. Let J denote the set of all elements of R which are divisible by at least one term of this sequence. Then J is an ideal. For if $a_j | b$ and $a_k | c$, where $j \leq k$, then also $a_k | b$ and hence $a_k | bx + cy$ for all $x, y \in R$. If J were generated by a single element a, we would have $a | a_n$ for every n. On the other hand, since $a \in J$, $a_N | a$ for some N. Hence $a_N | a_{N+1}$. Since a_{N+1} is a proper divisor of a_N, this is a contradiction. Thus R is not a principal ideal domain. □

It follows from the remarks at the end of Section 1 that a principal ideal domain is factorial, i.e. any element which is neither zero nor a unit can be represented as a product of finitely many irreducibles and the representation is essentially unique.

In the next section we will show that the ring $K[t]$ of all polynomials in one indeterminate t with coefficients from an arbitrary field K is a principal ideal domain.

It may be shown that the ring of all algebraic integers is a Bézout domain, and likewise the ring of all functions which are holomorphic in a nonempty connected open subset G of the complex plane $\mathbb{C}$. However, neither is a principal ideal domain. In the former case there are no irreducibles, since any algebraic integer a has the factorization $a = \sqrt{a} \cdot \sqrt{a}$. In the latter case $z - \zeta$ is an irreducible for any $\zeta \in G$, but the chain condition is violated. For example, take

$$a_n(z) = f(z)/(z - \zeta_1) \cdots (z - \zeta_n),$$

where $f(z)$ is a non-identically vanishing function which is holomorphic in G and has infinitely many zeros $\zeta_1, \zeta_2, \ldots$ in G.

3 Polynomials

In this section we study the most important example of a principal ideal domain other than $\mathbb{Z}$, namely the ring $K[t]$ of all polynomials in t with coefficients from an arbitrary field K (e.g., $K = \mathbb{Q}$ or $\mathbb{C}$).

The attitude adopted towards polynomials in algebra is different from that adopted in analysis. In analysis we regard 't' as a variable which can take different values; in algebra we regard 't' simply as a symbol, an 'indeterminate', on which we can perform various algebraic operations. Since the concept of function is so pervasive, the algebraic approach often seems mysterious at first sight and it seems worthwhile taking the time to give a precise meaning to an 'indeterminate'.

Let R be an integral domain (e.g., $R = \mathbb{Z}$ or $\mathbb{Q}$). A *polynomial* with coefficients from R is defined to be a sequence $f = (a_0, a_1, a_2, \ldots)$ of elements of R in which at most finitely many terms are nonzero. The sum and product of two polynomials

$$f = (a_0, a_1, a_2, \ldots), \quad g = (b_0, b_1, b_2, \ldots)$$

are defined by

$$\begin{aligned} f + g &= (a_0 + b_0, a_1 + b_1, a_2 + b_2, \ldots), \\ fg &= (a_0 b_0, a_0 b_1 + a_1 b_0, a_0 b_2 + a_1 b_1 + a_2 b_0, \ldots). \end{aligned}$$

It is easily verified that these are again polynomials and that the set $R[t]$ of all polynomials with coefficients from R is a commutative ring with $O = (0, 0, 0, \ldots)$ as zero element. (By dropping the requirement that at most finitely many terms are nonzero, we obtain the ring $R[[t]]$ of all *formal power series* with coefficients from R.)

We define the *degree* $\partial(f)$ of a polynomial $f = (a_0, a_1, a_2, \ldots) \neq O$ to be the greatest integer n for which $a_n \neq 0$ and we put

$$|f| = 2^{\partial(f)}, \quad |O| = 0.$$

It is easily verified that, for all polynomials f, g,

$$|f+g| \le \max\{|f|, |g|\}, \quad |fg| = |f||g|.$$

Since $|f| \ge 0$, with equality if and only if $f = O$, the last property implies that $R[t]$ *is an integral domain*. Thus we can define divisibility in $R[t]$, as explained in Section 1.

The set of all polynomials of the form $(a_0, 0, 0, \ldots)$ is a subdomain isomorphic to R. By identifying this set with R, we may regard R as embedded in $R[t]$. The only units in $R[t]$ are the units in R, since $1 = ef$ implies $1 = |e||f|$ and hence $|e| = 1$.

If we put $t = (0, 1, 0, 0, \ldots)$, then

$$t^2 = tt = (0, 0, 1, 0, \ldots), \quad t^3 = tt^2 = (0, 0, 0, 1, \ldots), \ldots.$$

Hence if the polynomial $f = (a_0, a_1, a_2, \ldots)$ has degree n, then it can be uniquely expressed in the form

$$f = a_0 + a_1 t + \cdots + a_n t^n \quad (a_n \ne 0).$$

We refer to the elements $a_0, a_1, \ldots, a_n$ of R as the *coefficients* of f. In particular, a_0 is the *constant* coefficient and a_n the *highest* coefficient. We say that f is *monic* if its highest coefficient $a_n = 1$.

If also

$$g = b_0 + b_1 t + \cdots + b_m t^m \quad (b_m \ne 0),$$

then the sum and product assume their familiar forms:

$$f + g = (a_0 + b_0) + (a_1 + b_1)t + (a_2 + b_2)t^2 + \cdots,$$
$$fg = a_0 b_0 + (a_0 b_1 + a_1 b_0)t + (a_0 b_2 + a_1 b_1 + a_2 b_0)t^2 + \cdots.$$

Suppose now that $R = K$ is a field, and let

$$f = a_0 + a_1 t + \cdots + a_n t^n \ (a_n \ne 0),$$
$$g = b_0 + b_1 t + \cdots + b_m t^m \ (b_m \ne 0)$$

be any two nonzero elements of $K[t]$. If $|g| < |f|$, i.e. if $m < n$, then $g = qf + r$, with $q = O$ and $r = g$. Suppose on the other hand that $|f| \le |g|$. Then

$$g = a_n^{-1} b_m t^{m-n} f + g^\dagger,$$

where $g^\dagger \in K[t]$ and $|g^\dagger| < |g|$. If $|f| \le |g^\dagger|$, the process can be repeated with $g^\dagger$ in place of g. Continuing in this way, we obtain $q, r \in K[t]$ such that

$$g = qf + r, \quad |r| < |f|.$$

Moreover, q and r are uniquely determined, since if also

$$g = q_1 f + r_1, \quad |r_1| < |f|,$$

then

$$(q - q_1)f = r_1 - r, \quad |r_1 - r| < |f|,$$

which is only possible if $q = q_1$.

Ideals in $K[t]$ can be defined in the same way as for $\mathbb{Z}$ and the proof of Lemma 9 remains valid. Thus $K[t]$ is a principal ideal domain and, *a fortiori*, a GCD domain.

The Euclidean algorithm can also be applied in $K[t]$ in the same way as for $\mathbb{Z}$ and again, from the sequence of polynomials $f_0, f_1, \ldots, f_N$ which it provides to determine the greatest common divisor f_N of f_0 and f_1 we can obtain polynomials u_k, v_k such that

$$f_k = f_1 u_k + f_0 v_k \quad (0 \leq k \leq N).$$

We can actually say more for polynomials than for integers, since if

$$f_{k-1} = q_k f_k + f_{k+1}, \quad |f_{k+1}| < |f_k|,$$

then $|f_{k-1}| = |q_k||f_k|$ and hence, by induction,

$$|f_{k-1}||u_k| = |f_0|, |f_{k-1}||v_k| = |f_1| \quad (1 < k \leq N).$$

It may be noted in passing that the Euclidean algorithm can also be applied in the ring $K[t, t^{-1}]$ of *Laurent polynomials*. A Laurent polynomial $f \neq O$, with coefficients from the field K, has the form

$$f = a_m t^m + a_{m+1} t^{m+1} + \cdots + a_n t^n,$$

where $m, n \in \mathbb{Z}$ with $m \leq n$ and $a_j \in K$ with $a_m a_n \neq 0$. Thus we can write $f = t^m f_0$, where $f_0 \in K[t]$. Put

$$|f| = 2^{n-m}, \quad |O| = 0;$$

then the division algorithm for ordinary polynomials implies one for Laurent polynomials: for any $f, g \in K[t, t^{-1}]$ with $f \neq O$, there exist $q, r \in K[t, t^{-1}]$ such that $g = qf + r, |r| < |f|$.

We return now to ordinary polynomials. The general definition for integral domains in Section 1 means, in the present case, that a polynomial $p \in K[t]$ is *irreducible* if it has positive degree and if every proper divisor has degree zero.

It follows that any polynomial of degree 1 is irreducible. However, there may exist also irreducible polynomials of higher degree. For example, we will show shortly that the polynomial $t^2 - 2$ is irreducible in $\mathbb{Q}[t]$. For $K = \mathbb{C}$, however, every irreducible polynomial has degree 1, by the fundamental theorem of algebra (Theorem I.30) and Proposition 14 below. It follows that, for $K = \mathbb{R}$, every irreducible polynomial has degree 1 or 2. (For if a real polynomial $f(t)$ has a root $\alpha \in \mathbb{C}\backslash\mathbb{R}$, its conjugate $\bar{\alpha}$ is also a root and $f(t)$ has the real irreducible factor $(t - \alpha)(t - \bar{\alpha})$.)

It is obvious that the chain condition (#) of Section 1 holds in the integral domain $K[t]$, since if g is a proper divisor of f, then $|g| < |f|$. It follows that any polynomial of positive degree can be represented as a product of finitely many irreducible polynomials and that the representation is essentially unique.

We now consider the connection between polynomials in the sense of algebra (polynomial forms) and polynomials in the sense of analysis (polynomial functions). Let K be a field and $f \in K[t]$:

$$f = a_0 + a_1 t + \cdots + a_n t^n.$$

If we replace 't' by $c \in K$ we obtain an element of K, which we denote by $f(c)$:

$$f(c) = a_0 + a_1 c + \cdots + a_n c^n.$$

A rapid procedure ('Horner's rule') for calculating $f(c)$ is to use the recurrence relations

$$f_0 = a_n, \ f_j = f_{j-1} c + a_{n-j} \quad (j = 1, \ldots, n).$$

It is readily shown by induction that

$$f_j = a_n c^j + a_{n-1} c^{j-1} + \cdots + a_{n-j},$$

and hence $f(c) = f_n$ is obtained with just n multiplications and n additions.

It is easily seen that $f = g + h$ implies $f(c) = g(c) + h(c)$, and $f = gh$ implies $f(c) = g(c)h(c)$. Thus the mapping $f \to f(c)$ is a 'homomorphism' of $K[t]$ into K. A simple consequence is the so-called *remainder theorem*:

Proposition 14 *Let K be a field and $c \in K$. If $f \in K[t]$, then*

$$f = (t - c)g + f(c),$$

for some $g \in K[t]$.

In particular, f is divisible by $t - c$ if and only if $f(c) = 0$.

Proof We already know that there exist $q, r \in K[t]$ such that

$$f = (t - c)q + r, \quad |r| \leq 1.$$

Thus $r \in K$ and the homomorphism properties imply that $f(c) = r$. □

We say that $c \in K$ is a *root* of the polynomial $f \in K[t]$ if $f(c) = 0$.

Proposition 15 *Let K be a field. If $f \in K[t]$ is a polynomial of degree $n \geq 0$, then f has at most n distinct roots in K.*

Proof If f is of degree 0, then $f = c$ is a nonzero element of K and f has no roots. Suppose now that $n \geq 1$ and the result holds for polynomials of degree less than n. If c is a root of f then, by Proposition 14, $f = (t - c)g$ for some $g \in K[t]$. Since g has degree $n - 1$, it has at most $n - 1$ roots. But every root of f distinct from c is a root of g. Hence f has at most n roots. □

We consider next properties of the integral domain $R[t]$, when R is an integral domain rather than a field (e.g., $R = \mathbb{Z}$). The famous Pythagorean proof that $\sqrt{2}$ is irrational is considerably generalized by the following result:

Proposition 16 *Let R be a GCD domain and K its field of fractions. Let*

$$f = a_0 + a_1 t + \cdots + a_n t^n$$

be a polynomial of degree $n > 0$ with coefficients $a_j \in R$ $(0 \leq j \leq n)$. If $c \in K$ is a root of f and $c = ab^{-1}$, where $a, b \in R$ and $(a, b) = 1$, then $b|a_n$ and $a|a_0$.

In particular, if f is monic, then $c \in R$.

Proof We have

$$a_0 b^n + a_1 a b^{n-1} + \cdots + a_{n-1} a^{n-1} b + a_n a^n = 0.$$

Hence $b|a_n a^n$ and $a|a_0 b^n$. Since $(a^n, b) = (a, b^n) = 1$, by Proposition 3(v), the result follows from Proposition 3(ii). □

The polynomial $t^2 - 2$ has no integer roots, since 0, 1, -1 are not roots and if $c \in \mathbb{Z}$ and $c \neq 0, 1, -1$, then $c^2 \geq 4$. Consequently, by Proposition 16, the polynomial $t^2 - 2$ also has no rational roots. It now follows from Proposition 14 that $t^2 - 2$ is irreducible in $\mathbb{Q}[t]$, since it has no divisors of degree 1.

Proposition 16 was known to Euler (1774) for the case $R = \mathbb{Z}$. In this case it shows that to obtain all rational roots of a polynomial with rational coefficients we need test only a finite number of possibilities, which can be explicitly enumerated. For example, if $z \in \mathbb{Z}$, the cubic polynomial $t^3 + zt + 1$ has no rational roots unless $z = 0$ or $z = -2$.

It was shown by Gauss (1801), again for the case $R = \mathbb{Z}$, that Proposition 16 may itself be considerably generalized. His result may be formulated in the following way:

Proposition 17 *Let $f, g \in R[t]$, where R is a GCD domain with field of fractions K. Then g divides f in $R[t]$ if and only if g divides f in $K[t]$ and the greatest common divisor of the coefficients of g divides the greatest common divisor of the coefficients of f.*

Proof For any polynomial $f \in R[t]$, let $c(f)$ denote the greatest common divisor of its coefficients. We say that f is *primitive* if $c(f) = 1$. We show first that the product $f = gh$ of two primitive polynomials g, h is again primitive.

Let

$$g = b_0 + b_1 t + \cdots, \quad h = c_0 + c_1 t + \cdots, \quad f = a_0 + a_1 t + \cdots,$$

and assume on the contrary that the coefficients a_i have a common divisor d which is not a unit. Then d does not divide all the coefficients b_j, nor all the coefficients c_k. Let b_m, c_n be the first coefficients of g, h which are not divisible by d. Then

$$a_{m+n} = \sum_{j+k=m+n} b_j c_k$$

and d divides every term on the right, except possibly $b_m c_n$. In fact, since $d|a_{m+n}$, d must also divide $b_m c_n$. Hence we cannot have both $(d, b_m) = 1$ and $(d, c_n) = 1$.

Consequently we can replace d by a proper divisor d', again not a unit, for which $m' + n' > m + n$. Since there exists a divisor d for which $m + n$ is a maximum, this yields a contradiction.

Now let f, g be polynomials in $R[t]$ such that g divides f in $K[t]$. Thus $f = gH$, where $H \in K[t]$. We can write $H = ab^{-1}h_0$, where a, b are coprime elements of R and h_0 is a primitive polynomial in $R[t]$. Also

$$f = c(f)f_0, \quad g = c(g)g_0,$$

where f_0, g_0 are primitive polynomials in $R[t]$. Hence

$$bc(f)f_0 = ac(g)g_0h_0.$$

Since g_0h_0 is primitive, it follows that

$$bc(f) = ac(g).$$

If $H \in R[t]$, then $b = 1$ and so $c(g)|c(f)$. On the other hand, if $c(g)|c(f)$, then $bc(f)/c(g) = a$. Since $(a, b) = 1$, this implies that $b = 1$ and $H \in R[t]$. □

Corollary 18 *If R is a GCD domain, then $R[t]$ is also a GCD domain. If, moreover, R is a factorial domain, then $R[t]$ is also a factorial domain.*

proof Let K denote the field of fractions of R. Since $K[t]$ is a GCD domain and $R[t] \subseteq K[t]$, $R[t]$ is certainly an integral domain. If $f, g \in R[t]$, then there exists a primitive polynomial $h_0 \in R[t]$ which is a greatest common divisor of f and g in $K[t]$. It follows from Proposition 17 that

$$h = (c(f), c(g))h_0$$

is a greatest common divisor of f and g in $R[t]$.

This proves the first statement of the corollary. It remains to show that if R also satisfies the chain condition (#), then $R[t]$ does likewise. But if $f_n \in R[t]$ and $f_{n+1}|f_n$ for every n, then f_n must be of constant degree for all large n. The second statement of the corollary now also follows from Proposition 17 and the chain condition in R. □

It follows by induction that in the statement of Corollary 18 we may replace $R[t]$ by the ring $R[t_1, \ldots, t_m]$ of all polynomials in finitely many indeterminates $t_1, \ldots, t_m$ with coefficients from R. In particular, if K is a field, then any polynomial $f \in K[t_1, \ldots, t_m]$ such that $f \notin K$ can be represented as a product of finitely many irreducible polynomials and the representation is essentially unique.

It is now easy to give examples of GCD domains which are not Bézout domains. Let R be a GCD domain which is not a field (e.g., $R = \mathbb{Z}$). Then some $a_0 \in R$ is neither zero nor a unit. By Corollary 18, $R[t]$ is a GCD domain and, by Proposition 17, the greatest common divisor in $R[t]$ of the polynomials a_0 and t is 1. If there existed $g, h \in R[t]$ such that

$$a_0g + th = 1,$$

where $g = b_0 + b_1 t + \cdots$, then by equating constant coefficients we would obtain $a_0 b_0 = 1$, which is a contradiction. Thus $R[t]$ is not a Bézout domain.

As an application of the preceding results we show that if $a_1, \ldots, a_n$ are distinct integers, then the polynomial

$$f = \prod_{j=1}^{n} (t - a_j) - 1$$

is irreducible in $\mathbb{Q}[t]$. Assume, on the contrary, that $f = gh$, where $g, h \in \mathbb{Q}[t]$ and have positive degree. We may suppose without loss of generality that $g \in \mathbb{Z}[t]$ and that the greatest common divisor of the coefficients of g is 1. Since $f \in \mathbb{Z}[t]$, it then follows from Proposition 17 that also $h \in \mathbb{Z}[t]$. Thus $g(a_j)$ and $h(a_j)$ are integers for every j. Since $g(a_j)h(a_j) = -1$, it follows that $g(a_j) = -h(a_j)$. Thus the polynomial $g + h$ has the distinct roots $a_1, \ldots, a_n$. Since $g + h$ has degree less than n, this is possible only if $g + h = O$. Hence $f = -g^2$. But, since the highest coefficient of f is 1, this is a contradiction.

In general, it is not an easy matter to determine if a polynomial with rational coefficients is irreducible in $\mathbb{Q}[t]$. However, the following *irreducibility criterion*, due to Eisenstein (1850), is sometimes useful:

Proposition 19 *If*

$$f(t) = a_0 + a_1 t + \cdots + a_{n-1} t^{n-1} + t^n$$

is a monic polynomial of degree n with integer coefficients such that $a_0, a_1, \ldots, a_{n-1}$ are all divisible by some prime p, but a_0 is not divisible by p^2, then f is irreducible in $\mathbb{Q}[t]$.

Proof Assume on the contrary that f is reducible. Then there exist polynomials $g(t), h(t)$ of positive degrees l, m with *integer* coefficients such that $f = gh$. If

$$g(t) = b_0 + b_1 t + \cdots + b_l t^l,$$
$$h(t) = c_0 + c_1 t + \cdots + c_m t^m,$$

then $a_0 = b_0 c_0$. The hypotheses imply that exactly one of b_0, c_0 is divisible by p. Without loss of generality, assume it to be b_0. Since p divides $a_1 = b_0 c_1 + b_1 c_0$, it follows that $p | b_1$. Since p divides $a_2 = b_0 c_2 + b_1 c_1 + b_2 c_0$, it now follows that $p | b_2$. Proceeding in this way, we see that p divides b_j for every $j \leq l$. But, since $b_l c_m = 1$, this yields a contradiction. □

It follows from Proposition 19 that, for any prime p, the p-th *cyclotomic polynomial*

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

is irreducible in $\mathbb{Q}[x]$. For $\Phi_p(x) = (x^p - 1)/(x - 1)$ and, if we put $x = 1 + t$, the transformed polynomial

$$\{(1+t)^p - 1\}/t = t^{p-1} + {}^pC_{p-1}t^{p-2} + \cdots + {}^pC_2 t + p$$

satisfies the hypotheses of Proposition 19.

For any field K, we define the *formal derivative* of a polynomial $f \in K[t]$,

$$f = a_0 + a_1 t + \cdots + a_n t^n,$$

to be the polynomial

$$f' = a_1 + 2a_2 t + \cdots + n a_n t^{n-1}.$$

If the field K is of *characteristic* 0 (see Chapter I, §8), then $\partial(f') = \partial(f) - 1$.

Formal derivatives share the following properties with the derivatives of real analysis:

(i) $(f+g)' = f' + g'$;
(ii) $(cf)' = cf'$ *for any* $c \in K$;
(iii) $(fg)' = f'g + fg'$;
(iv) $(f^k)' = kf^{k-1}f'$ *for any* $k \in \mathbb{N}$.

The first two properties are easily established and the last two properties then need only be verified for monomials $f = t^m$, $g = t^n$.

We can use formal derivatives to determine when a polynomial is *square-free*:

Proposition 20 *Let f be a polynomial of positive degree with coefficients from a field K. If f is relatively prime to its formal derivative f', then f is a product of irreducible polynomials, no two of which differ by a constant factor. Conversely, if f is such a product and if K has characteristic 0, then f is relatively prime to f'.*

Proof If $f = g^2 h$ for some polynomials $g, h \in K[t]$ with $\partial(g) > 0$ then, by the rules above,

$$f' = 2gg'h + g^2 h'.$$

Hence $g \mid f'$ and f, f' are not relatively prime.

On the other hand, if $f = p_1 \cdots p_m$ is a product of essentially distinct irreducible polynomials p_j, then

$$f' = p_1' p_2 \cdots p_m + p_1 p_2' p_3 \cdots p_m + \cdots + p_1 \cdots p_{m-1} p_m'.$$

If the field K has characteristic 0, then p_1' is of lower degree than p_1 and is not the zero polynomial. Thus the first term on the right is not divisible by p_1, but all the other terms are. Therefore $p_1 \nmid f'$, and hence $(f', p_1) = 1$. Similarly, $(f', p_j) = 1$ for $1 < j \leq m$. Since essentially distinct irreducible polynomials are relatively prime, it follows that $(f', f) = 1$. □

For example, it follows from Proposition 20 that the polynomial $t^n - 1 \in K[t]$ is square-free if the characteristic of the field K does not divide the positive integer n.

4 Euclidean Domains

An integral domain R is said to be *Euclidean* if it possesses a Euclidean algorithm, i.e. if there exists a map $\delta\colon R \to \mathbb{N} \cup \{0\}$ such that, for any $a, b \in R$ with $a \neq 0$, there exist $q, r \in R$ with the properties

$$b = qa + r, \quad \delta(r) < \delta(a).$$

It follows that $\delta(a) > \delta(0)$ for any $a \neq 0$. For there exist $q_1, a_1 \in R$ such that

$$0 = q_1 a + a_1, \quad \delta(a_1) < \delta(a),$$

and if $a_n \neq 0$ there exist $q_{n+1}, a_{n+1} \in R$ such that

$$0 = q_{n+1} a_n + a_{n+1}, \quad \delta(a_{n+1}) < \delta(a_n).$$

Repeatedly applying this process, we must arrive at $a_N = 0$ for some N, since the sequence $\{\delta(a_n)\}$ cannot decrease forever, and we then have $\delta(0) = \delta(a_N) < \cdots < \delta(a_1) < \delta(a)$.

By replacing δ by $\delta - \delta(0)$ *we may, and will, assume that* $\delta(0) = 0$, $\delta(a) > 0$ if $a \neq 0$.

Since the proof of Lemma 9 remains valid if $\mathbb{Z}$ is replaced by R and $|a|$ by $\delta(a)$, *any Euclidean domain is a principal ideal domain*.

The *polynomial ring* $K[t]$ is a Euclidean domain with $\delta(a) = |a| = 2^{\partial(a)}$. Polynomial rings are characterized among all Euclidean domains by the following result:

Proposition 21 *For a Euclidean domain R, the following conditions are equivalent:*

(i) *for any* $a, b \in R$ *with* $a \neq 0$, *there exist unique* $q, r \in R$ *such that* $b = qa + r$, $\delta(r) < \delta(a)$;

(ii) *for any* $a, b, c \in R$ *with* $c \neq 0$,

$$\delta(a + b) \leq \max\{\delta(a), \delta(b)\}, \quad \delta(a) \leq \delta(ac).$$

Moreover, if one or other of these two conditions holds, then either R is a field and $\delta(a) = \delta(1)$ *for every* $a \neq 0$, *or* $R = K[t]$ *for some field* K *and* δ *is an increasing function of* $|\ |$.

Proof Suppose first that (i) holds. If $a \neq 0$, $c \neq 0$, then from $0 = 0a - 0 = ca - ac$, we obtain $\delta(ac) \geq \delta(a)$, and this holds also if $a = 0$. If we take $c = -1$ and replace a by $-a$, we get $\delta(-a) = \delta(a)$. Since $b = 0(a + b) + b = 1(a + b) + (-a)$, it follows that either $\delta(b) \geq \delta(a + b)$ or $\delta(a) \geq \delta(a + b)$. Thus (i) $\Rightarrow$ (ii).

Suppose next that (ii) holds. Assume that, for some $a, b \in R$ with $a \neq 0$, there exist pairs q, r and q', r' such that

$$b = qa + r = q'a + r', \quad \max\{\delta(r), \delta(r')\} < \delta(a).$$

From (ii) we obtain first $\delta(-r) = \delta(r)$ and then $\delta(r' - r) \leq \max\{\delta(r), \delta(r')\} < \delta(a)$. Since $r' - r = a(q - q')$, this implies $q - q' = 0$ and hence $r' - r = 0$. Thus (ii) $\Rightarrow$ (i).

Suppose now that (i) and (ii) both hold. Then $\delta(1) \leq \delta(a)$ for any $a \neq 0$, since $a = 1a$. Furthermore, $\delta(a) = \delta(ae)$ for any unit e, since

$$\delta(a) \leq \delta(ae) \leq \delta(aee^{-1}) = \delta(a).$$

On the other hand, $\delta(a) = \delta(ae)$ for some $a \neq 0$ implies that e is a unit. For from

$$a = qae + r, \quad \delta(r) < \delta(ae),$$

we obtain $r = (1-qe)a$, $\delta(r) < \delta(a)$, and hence $1-qe = 0$. In particular, $\delta(e) = \delta(1)$ if and only if e is a unit.

The set K of all $a \in R$ such that $\delta(a) \leq \delta(1)$ thus consists of 0 and all units of R. Since $a, b \in K$ implies $a - b \in K$, it follows that K is a field. We assume that $K \neq R$, since otherwise we have the first alternative of the proposition.

Choose $x \in R \backslash K$ so that

$$\delta(x) = \min_{a \in R \backslash K} \delta(a).$$

For any $a \in R \backslash K$, there exist $q_0, r_0 \in R$ such that

$$a = q_0 x + r_0, \quad \delta(r_0) < \delta(x),$$

i.e. $r_0 \in K$. Then $\delta(q_0) < \delta(q_0 x) = \delta(a - r_0) \leq \delta(a)$. If $\delta(q_0) \geq \delta(x)$, i.e. if $q_0 \in R \backslash K$, then in the same way there exist $q_1, r_1 \in R$ such that

$$q_0 = q_1 x + r_1, r_1 \in K, \quad \delta(q_1) < \delta(q_0).$$

After finitely many repetitions of this process we must arrive at some $q_{n-1} \in K$. Putting $r_n = q_{n-1}$, we obtain

$$a = r_n x^n + r_{n-1} x^{n-1} + \cdots + r_0,$$

where $r_0, \ldots, r_n \in K$ and $r_n \neq 0$. Since $\delta(r_j x^j) = \delta(x^j)$ if $r_j \neq 0$ and $\delta(x^j) < \delta(x^{j+1})$ for every j, it follows that $\delta(a) = \delta(x^n)$. Since the representation $a = qx^n + r$ with $\delta(r) < \delta(x^n)$ is unique, it follows that $r_0, \ldots, r_n$ are uniquely determined by a. Define a map $\psi : R \to K[t]$ by

$$\psi(r_n x^n + r_{n-1} x^{n-1} + \cdots + r_0) = r_n t^n + r_{n-1} t^{n-1} + \cdots + r_0.$$

Then ψ is a bijection and actually an isomorphism, since it preserves sums and products. Furthermore $\delta(a) >, =,$ or $< \delta(b)$ according as $|\psi(a)| >, =,$ or $< |\psi(b)|$. □

Some significant examples of principal ideal domains are provided by quadratic fields, which will be studied in Chapter III. Any quadratic number field has the form $\mathbb{Q}(\sqrt{d})$, where $d \in \mathbb{Z}$ is square-free and $d \neq 1$. The set $\mathcal{O}_d$ of all algebraic integers in $\mathbb{Q}(\sqrt{d})$ is an integral domain. In the equivalent language of binary quadratic forms, it was known to Gauss that $\mathcal{O}_d$ is a principal ideal domain for nine negative values of d, namely

$$d = -1, -2, -3, -7, -11, -19, -43, -67, -163.$$

Heilbronn and Linfoot (1934) showed that there was at most one additional negative value of d for which $\mathcal{O}_d$ is a principal ideal domain. Stark (1967) proved that this additional value does not in fact exist, and soon afterwards it was observed that a gap in a previous proof by Heegner (1952) could be filled without difficulty. It is conjectured that $\mathcal{O}_d$ is a principal ideal domain for infinitely many positive values of d, but this remains unproved.

Much work has been done on determining for which quadratic number fields $\mathbb{Q}(\sqrt{d})$ the ring of integers $\mathcal{O}_d$ is a Euclidean domain. Although we regard being Euclidean more as a useful property than as an important concept, we report here the results which have been obtained for their intrinsic interest.

The ring $\mathcal{O}_d$ is said to be *norm-Euclidean* if it is Euclidean when one takes $\delta(a)$ to be the absolute value of the *norm* of a. It has been shown that $\mathcal{O}_d$ is norm-Euclidean for precisely the following values of d:

$$d = -11, -7, -3, -2, -1, 2, 3, 5, 6, 7, 11, 13, 17, 19, 21, 29, 33, 37, 41, 57, 73.$$

It is known that, for $d < 0$, $\mathcal{O}_d$ is Euclidean only if it is norm-Euclidean. Comparing the two lists, we see that for $d = -19, -43, -67, -163$, $\mathcal{O}_d$ is a principal ideal domain, but not a Euclidean domain. On the other hand it is also known that, for $d = 69$, $\mathcal{O}_d$ is Euclidean but not norm-Euclidean.

5 Congruences

The invention of a new notation often enables one to replace a long, involved argument by simple and mechanical algebraic operations. This is well illustrated by the congruence notation.

Two integers a and b are said to be *congruent modulo* a third integer m if m divides $a - b$, and this is denoted by $a \equiv b \bmod m$. For example,

$$13 \equiv 4 \bmod 3, \quad 13 \equiv -7 \bmod 5, \quad 19 \equiv 7 \bmod 4.$$

The notation is a modification by Gauss of the notation $a = b \bmod m$ used by Legendre, as Gauss explicitly acknowledged (*D.A.*, §2). (If a and b are not congruent modulo m, we write $a \not\equiv b \bmod m$.) Congruence has, in fact, many properties in common with equality:

(C1) $a \equiv a \bmod m$ *for all* a, m; (reflexive law)
(C2) *if* $a \equiv b \bmod m$, *then* $b \equiv a \bmod m$; (symmetric law)
(C3) *if* $a \equiv b$ *and* $b \equiv c \bmod m$, *then* $a \equiv c \bmod m$; (transitive law)
(C4) *if* $a \equiv a'$ *and* $b \equiv b' \bmod m$, *then* $a + b \equiv a' + b'$ *and* $ab \equiv a'b' \bmod m$. (replacement laws)

The proofs of these properties are very simple. For any a, m we have $a - a = 0 = m \cdot 0$. If m divides $a - b$, then it also divides $b - a = -(a - b)$. If m divides both $a - b$ and $b - c$, then it also divides $(a - b) + (b - c) = a - c$. Finally, if m divides both $a - a'$ and $b - b'$, then it also divides $(a - a') + (b - b') = (a + b) - (a' + b')$ and $(a - a')b + a'(b - b') = ab - a'b'$.

The properties **(C1)–(C3)** state that congruence mod m is an *equivalence relation*. Since $a = b$ implies $a \equiv b \bmod m$, it is a coarsening of the equivalence relation of

equality (but coincides with it if $m = 0$). The corresponding equivalence classes are called *residue classes*. The set $\mathbb{Z}$ with equality replaced by congruence mod m will be denoted by $\mathbb{Z}_{(m)}$. If $m > 0$, $\mathbb{Z}_{(m)}$ has cardinality m, since an arbitrary integer a can be uniquely represented in the form $a = qm + r$, where $r \in \{0, 1, \ldots, m-1\}$ and $q \in \mathbb{Z}$. The particular r which represents a given $a \in \mathbb{Z}$ is referred to as the *least non-negative residue* of $a \bmod m$.

The replacement laws imply that the associative, commutative and distributive laws for addition and multiplication are inherited from $\mathbb{Z}$ by $\mathbb{Z}_{(m)}$. Hence $\mathbb{Z}_{(m)}$ is a commutative ring, with 0 as an identity element for addition and 1 as an identity element for multiplication. However, $\mathbb{Z}_{(m)}$ is not an integral domain if m is composite, since if $m = m'm''$ with $1 < m' < m$, then

$$m'm'' \equiv 0, \text{ but } m' \not\equiv 0, m'' \not\equiv 0 \bmod m.$$

On the other hand, if $ab \equiv ac \bmod m$ and $(a, m) = 1$, then $b \equiv c \bmod m$, by Proposition 3(ii). Thus factors which are relatively prime to the modulus can be cancelled.

In algebraic terms, $\mathbb{Z}_{(m)}$ is the *quotient ring* $\mathbb{Z}/m\mathbb{Z}$ of $\mathbb{Z}$ with respect to the ideal $m\mathbb{Z}$ generated by m, and the elements of $\mathbb{Z}_{(m)}$ are the *cosets* of this ideal. For convenience, rather than necessity, we suppose from now on that $m > 1$.

Congruences enter implicitly into many everyday problems. For example, the ring $\mathbb{Z}_{(2)}$ contains two distinct elements, 0 and 1, with the addition and multiplication tables

$$0 + 0 = 1 + 1 = 0, 0 + 1 = 1 + 0 = 1,$$
$$0 \cdot 0 = 0 \cdot 1 = 1 \cdot 0 = 0, 1 \cdot 1 = 1.$$

This is the arithmetic of *odds* (1) and *evens* (0), which is used by electronic computers.

Again, to determine the day of the week on which one was born, from the date and day of the week today, is an easy calculation in the arithmetic of $\mathbb{Z}_{(7)}$ (remembering that $366 \equiv 2 \bmod 7$).

The well-known tests for divisibility of an integer by 3 or 9 are easily derived by means of congruences. Let the positive integer a have the decimal representation

$$a = a_0 + a_1 10 + \cdots + a_n 10^n,$$

where $a_0, a_1, \ldots, a_n \in \{0, 1, \ldots, 9\}$. Since $10 \equiv 1 \bmod m$, where $m = 3$ or 9, the replacement laws imply that $10^k \equiv 1 \bmod m$ for any positive integer k and hence

$$a \equiv a_0 + a_1 + \cdots + a_n \bmod m.$$

Thus a is divisible by 3 or 9 if and only if the sum of its digits is so divisible.

This can be used to check the accuracy of arithmetical calculations. Any equation involving only additions and multiplications must remain valid when equality is replaced by congruence mod m. For example, suppose we wish to check if

$$7714 \times 3036 = 23{,}419{,}804.$$

Taking congruences mod 9, we have on the left side $19 \times 12 \equiv 1 \times 3 \equiv 3$ and on the right side $5 + 14 + 12 \equiv 5 + 5 + 3 \equiv 4$. Since $4 \not\equiv 3 \bmod 9$, the original equation is incorrect (the 8 should be a 7).

Since the distinct squares in $\mathbb{Z}_{(4)}$ are 0 and 1, it follows that an integer $a \equiv 3 \bmod 4$ cannot be represented as the sum of two squares of integers. Similarly, since the distinct squares in $\mathbb{Z}_{(8)}$ are 0,1,4, an integer $a \equiv 7 \bmod 8$ cannot be represented as the sum of three squares of integers.

The oldest known work on number theory is a Babylonian cuneiform text, from at least as early as 1600 B.C., which contains a list of right-angled triangles whose side lengths are all exact multiples of the unit length. By Pythagoras' theorem, the problem is to find positive integers x, y, z such that

$$x^2 + y^2 = z^2.$$

For example, 3, 4, 5 and 5, 12, 13 are solutions. The number of solutions listed suggests that the Babylonians not only knew the theorem of Pythagoras, but also had some rule for finding such *Pythagorean triples*. There are in fact infinitely many, and a rule for finding them all is given by Euclid in his *Elements* (Book X, Lemma 1 following Proposition 28). This rule will now be derived.

We may assume that x and y are relatively prime since, if x, y, z is a Pythagorean triple for which x and y have greatest common divisor d, then $d^2|z^2$ and hence $d|z$, so that $x/d, y/d, z/d$ is also a Pythagorean triple. If x and y are relatively prime, then they are not both even and without loss of generality we may assume that x is odd. If y were also odd, we would have

$$z^2 = x^2 + y^2 \equiv 1 + 1 \equiv 2 \bmod 4,$$

which is impossible. Hence y is even and z is odd. Then 2 is a common divisor of $z + x$ and $z - x$, and is actually their greatest common divisor, since $(x, y) = 1$ implies $(x, z) = 1$. Since

$$(y/2)^2 = (z + x)/2 \cdot (z - x)/2$$

and the two factors on the right are relatively prime, they are also squares:

$$(z + x)/2 = a^2, \quad (z - x)/2 = b^2,$$

where $a > b > 0$ and $(a, b) = 1$. Then

$$x = a^2 - b^2, \quad y = 2ab, \quad z = a^2 + b^2.$$

Moreover a and b cannot both be odd, since z is odd.

Conversely, if x, y, z are defined by these formulas, where a and b are relatively prime positive integers with $a > b$ and either a or b even, then x, y, z is a Pythagorean triple. Moreover x is odd, since z is odd and y even, and it is easily verified that $(x, y) = 1$. For given x and z, a^2 and b^2 are uniquely determined, and hence a and b are also. Thus different couples a, b give different solutions x, y, z.

To return to congruences, we now consider the structure of the ring $\mathbb{Z}_{(m)}$. If $a \equiv a' \bmod m$ and $(a, m) = 1$, then also $(a', m) = 1$. Hence we may speak of an element of $\mathbb{Z}_{(m)}$ as being relatively prime to m. The set of all elements of $\mathbb{Z}_{(m)}$ which are relatively prime to m will be denoted by $\mathbb{Z}_{(m)}^{\times}$. If a is a *unit* of the ring $\mathbb{Z}_{(m)}$, then clearly $a \in \mathbb{Z}_{(m)}^{\times}$. The following proposition shows that, conversely, if $a \in \mathbb{Z}_{(m)}^{\times}$, then a is a unit of the ring $\mathbb{Z}_{(m)}$.

Proposition 22 *The set $\mathbb{Z}_{(m)}^{\times}$ is a commutative group under multiplication.*

Proof By Proposition 3(iv), $\mathbb{Z}_{(m)}^{\times}$ is closed under multiplication. Since multiplication is associative and commutative, it only remains to show that any $a \in \mathbb{Z}_{(m)}^{\times}$ has an inverse $a^{-1} \in \mathbb{Z}_{(m)}^{\times}$.

The elements of $\mathbb{Z}_{(m)}^{\times}$ may be taken to be the positive integers $c_1, \ldots, c_h$ which are less than m and relatively prime to m, and we may choose the notation so that $c_1 = 1$. Since $ac_j \equiv ac_k \bmod m$ implies $c_j \equiv c_k \bmod m$, the elements $ac_1, \ldots, ac_h$ are distinct elements of $\mathbb{Z}_{(m)}^{\times}$ and hence are a permutation of $c_1, \ldots, c_h$. In particular, $ac_i \equiv c_1 \bmod m$ for one and only one value of i. (The existence of inverses also follows from the Bézout identity $au + mv = 1$, since this implies $au \equiv 1 \bmod m$. Hence the Euclidean algorithm provides a way of calculating a^{-1}.) □

Corollary 23 *If p is a prime, then $\mathbb{Z}_{(p)}$ is a finite field with p elements.*

Proof We already know that $\mathbb{Z}_{(p)}$ is a commutative ring, whose distinct elements are represented by the integers $0, 1, \ldots, p-1$. Since p is a prime, $\mathbb{Z}_{(p)}^{\times}$ consists of all nonzero elements of $\mathbb{Z}_{(p)}$. Since $\mathbb{Z}_{(p)}^{\times}$ is a multiplicative group, by Proposition 22, it follows that $\mathbb{Z}_{(p)}$ is a field. □

The finite field $\mathbb{Z}_{(p)}$ will be denoted from now on by the more usual notation $\mathbb{F}_p$. Corollary 23, in conjunction with Proposition 15, implies that if p is a prime and f a polynomial of degree $n \geq 1$, then the congruence

$$f(x) \equiv 0 \bmod p$$

has at most n mutually incongruent solutions mod p. This is no longer true if the modulus is not a prime. For example, the congruence $x^2 - 1 \equiv 0 \bmod 8$ has the distinct solutions $x \equiv 1, 3, 5, 7 \bmod 8$.

The *order* of the group $\mathbb{Z}_{(m)}^{\times}$, i.e. the number of positive integers less than m and relatively prime to m, is traditionally denoted by $\varphi(m)$, with the convention that $\varphi(1) = 1$. For example, if p is a prime, then $\varphi(p) = p - 1$. More generally, for any positive integer k,

$$\varphi(p^k) = p^k - p^{k-1},$$

since the elements of $\mathbb{Z}_{(p^k)}$ which are not in $\mathbb{Z}_{(p^k)}^{\times}$ are the multiples jp with $0 \leq j < p^{k-1}$. By Proposition 4, if $m = m'm''$, where $(m', m'') = 1$, then $\varphi(m) = \varphi(m')\varphi(m'')$. Together with what we have just proved, this implies that if an arbitrary positive integer m has the factorization

$$m = p_1^{k_1} \cdots p_s^{k_s}$$

as a product of positive powers of distinct primes, then

$$\varphi(m) = p_1^{k_1-1}(p_1 - 1) \cdots p_s^{k_s-1}(p_s - 1).$$

In other words,

$$\varphi(m) = m \prod_{p|m} (1 - 1/p).$$

The function $\varphi(m)$ was first studied by Euler and is known as Euler's *phi*-function (or 'totient' function), although it was Gauss who decided on the letter φ. Gauss (*D.A.*, §39) also established the following property:

Proposition 24 *For any positive integer n,*

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

where the summation is over all positive divisors d of n.

Proof Let d be a positive divisor of n and let S_d denote the set of all positive integers $m \leq n$ such that $(m, n) = d$. Since $(m, n) = d$ if and only if $(m/d, n/d) = 1$, the cardinality of S_d is $\varphi(n/d)$. Moreover every positive integer $m \leq n$ belongs to exactly one such set S_d. Hence

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

since n/d runs through the positive divisors of n at the same time as d. □

Much of the significance of Euler's function stems from the following property:

Proposition 25 *If m is a positive integer and a an integer relatively prime to m, then*

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

Proof Let $c_1, \ldots, c_h$, where $h = \varphi(m)$, be the distinct elements of $\mathbb{Z}^\times_{(m)}$. As we saw in the proof of Proposition 22, the elements $ac_1, \ldots, ac_h$ of $\mathbb{Z}^\times_{(m)}$ are just a permutation of $c_1, \ldots, c_h$. Forming their product, we obtain $a^h c_1 \cdots c_h \equiv c_1 \cdots c_h \bmod m$. Since the c's are relatively prime to m, they can be cancelled and we are left with $a^h \equiv 1 \bmod m$. □

Corollary 26 *If p is a prime and a an integer not divisible by p, then* $a^{p-1} \equiv 1 \bmod p$.

Corollary 26 was stated without proof by Fermat (1640) and is commonly known as 'Fermat's little theorem'. The first published proof was given by Euler (1736), who later (1760) proved the general Proposition 25.

Proposition 25 is actually a very special case of Lagrange's theorem that the order of a subgroup of a finite group divides the order of the whole group. In the present case the whole group is $\mathbb{Z}^\times_{(m)}$ and the subgroup is the cyclic group generated by a.

Euler gave also another proof of Corollary 26, which has its own interest. For any two integers a, b and any prime p we have, by the binomial theorem,

$$(a+b)^p = \sum_{k=0}^{p} {}^pC_k a^k b^{p-k},$$

where the binomial coefficients

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

are integers. Moreover p divides ${}^pC_k$ for $0 < k < p$, since p divides ${}^pC_k \cdot k!$ and is relatively prime to $k!$ It follows that

$$(a+b)^p \equiv a^p + b^p \bmod p.$$

In particular, $(a+1)^p \equiv a^p + 1 \bmod p$, from which we obtain by induction $a^p \equiv a \bmod p$ for every integer a. If p does not divide a, the factor a can be cancelled to give $a^{p-1} \equiv 1 \bmod p$.

The first part of the second proof actually shows that *in any commutative ring R, of prime characteristic p, the map $a \to a^p$ is a homomorphism*:

$$(a+b)^p = a^p + b^p, \quad (ab)^p = a^p b^p.$$

(As defined in §8 of Chapter I, R has *characteristic k* if k is the least positive integer such that the sum of k 1's is 0, and has *characteristic zero* if there is no such positive integer.) By way of illustration, we give one important application of this result.

We showed in §3 that, for any prime p, the polynomial

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

is irreducible in $\mathbb{Q}[x]$. The roots in $\mathbb{C}$ of $\Phi_p(x)$ are the p-th roots of unity, other than 1. By a quite different argument we now show that, for any positive integer n, the 'primitive' n-th roots of unity are the roots of a monic polynomial $\Phi_n(x)$ with integer coefficients which is irreducible in $\mathbb{Q}[x]$. The uniquely determined polynomial $\Phi_n(x)$ is called the n-th *cyclotomic polynomial*.

Let ζ be a *primitive* n-th root of unity, i.e. $\zeta^n = 1$ but $\zeta^k \neq 1$ for $0 < k < n$. It follows from Corollary 18 that ζ is a root of some monic irreducible polynomial $f(x) \in \mathbb{Z}[x]$ which divides $x^n - 1$. If p is a prime which does not divide n, then ζ^p is also a primitive n-th root of unity and, for the same reason, ζ^p is a root of some monic irreducible polynomial $g(x) \in \mathbb{Z}[x]$ which divides $x^n - 1$.

We show first that $g(x) = f(x)$. Assume on the contrary that $g(x) \neq f(x)$. Then

$$x^n - 1 = f(x)g(x)h(x)$$

for some $h(x) \in \mathbb{Z}[x]$. Since ζ is a root of $g(x^p)$, we also have

$$g(x^p) = f(x)k(x)$$

for some $k(x) \in \mathbb{Z}[x]$. If $\bar{f}(x), \ldots$ denotes the polynomial in $\mathbb{F}_p[x]$ obtained from $f(x), \ldots$ by reducing the coefficients mod p,

then

$$x^n - 1 = \bar{f}(x)\bar{g}(x)\bar{h}(x), \quad \bar{g}(x^p) = \bar{f}(x)\bar{k}(x).$$

But $\bar{g}(x^p) = \bar{g}(x)^p$, since $\mathbb{F}_p[x]$ is a ring of characteristic p and $a^p = a$ for every $a \in \mathbb{F}_p$. Hence any irreducible factor $\bar{e}(x)$ of $\bar{f}(x)$ in $\mathbb{F}_p[x]$ also divides $\bar{g}(x)$. Consequently $\bar{e}(x)^2$ divides $x^n - 1$ in $\mathbb{F}_p[x]$. But $x^n - 1$ is relatively prime to its formal derivative nx^{n-1}, since $p \nmid n$, and so is square-free. This is the desired contradiction.

By applying this repeatedly for the same or different primes p, we see that ζ^m is a root of $f(x)$ for any positive integer m less than n and relatively prime to n. If ω is any n-th root of unity, then $\omega = \zeta^k$ for a unique k such that $0 \le k < n$. If $(k, n) \ne 1$, then $\omega^d = 1$ for some proper divisor d of n (cf. Lemma 31 below). If such an ω were a root of $f(x)$, then $f(x)$ would divide $x^d - 1$, which is impossible since ζ is not a root of $x^d - 1$. Hence $f(x)$ does not depend on the original choice of primitive n-th root of unity, its roots being all the primitive n-th roots of unity. The polynomial $f(x)$ will now be denoted by $\Phi_n(x)$. Since $x^n - 1$ is square-free, we have

$$x^n - 1 = \prod_{d|n} \Phi_d(x).$$

This yields a new proof of Proposition 24, since $\Phi_d(x)$ has degree $\varphi(d)$.

As an application of Fermat's little theorem (Corollary 26) we now prove

Proposition 27 *If p is a prime, then $(p-1)! + 1$ is divisible by p.*

Proof Since $1! + 1 = 2$, we may suppose that the prime p is odd. By Corollary 26, the polynomial $f(t) = t^{p-1} - 1$ has the distinct roots $1, 2, \ldots, p-1$ in the field $\mathbb{F}_p$. But the polynomial $g(t) = (t-1)(t-2)\cdots(t-p+1)$ has the same roots. Since $f(t) - g(t)$ is a polynomial of degree less than $p-1$, it follows from Proposition 15 that $f(t) - g(t)$ is the zero polynomial. In particular, $f(t)$ and $g(t)$ have the same constant coefficient. Since $(-1)^{p-1} = 1$, this yields the result. □

Proposition 27 is known as *Wilson's theorem*, although the first published proof was given by Lagrange (1773). Lagrange observed also that $(n-1)! + 1$ is divisible by n *only* if n is prime. For suppose $n = n'n''$, where $1 < n', n'' < n$. If $n' \ne n''$, then both n' and n'' occur as factors in $(n-1)!$ and hence n divides $(n-1)!$ If $n' = n'' > 2$ then, since $n > 2n'$, both n' and $2n'$ occur as factors in $(n-1)!$ and again n divides $(n-1)!$ Finally, if $n = 4$, then n divides $(n-1)! + 2$.

As another application of Fermat's little theorem, we prove *Euler's criterion for quadratic residues*. If p is a prime and a an integer not divisible by p, we say that a is a *quadratic residue,* or *quadratic nonresidue,* of p according as there exists, or does not exist, an integer c such that $c^2 \equiv a \bmod p$. Thus a is a quadratic residue of p if and only if it is a square in $\mathbb{F}_p^\times$. Euler's criterion is the first statement of the following proposition:

Proposition 28 *If p is an odd prime and a an integer not divisible by p, then*

$$a^{(p-1)/2} \equiv 1 \text{ or } -1 \bmod p,$$

according as a is a quadratic residue or nonresidue of p.

Moreover, exactly half of the integers $1, 2, \ldots, p-1$ are quadratic residues of p.

Proof If a is a quadratic residue of p, then $a \equiv c^2 \bmod p$ for some integer c and hence, by Fermat's little theorem,

$$a^{(p-1)/2} \equiv c^{p-1} \equiv 1 \bmod p.$$

Since the polynomial $t^{(p-1)/2} - 1$ has at most $(p-1)/2$ roots in the field $\mathbb{F}_p$, it follows that there are at most $r := (p-1)/2$ distinct quadratic residues of p. On the other hand, no two of the integers $1^2, 2^2, \ldots, r^2$ are congruent mod p, since $u^2 \equiv v^2 \bmod p$ implies $u \equiv v$ or $u \equiv -v \bmod p$. Hence there are exactly $(p-1)/2$ distinct quadratic residues of p and, if b is a quadratic nonresidue of p, then $b^{(p-1)/2} \not\equiv 1 \bmod p$. Since $b^{p-1} \equiv 1 \bmod p$, and

$$b^{p-1} - 1 = (b^{(p-1)/2} - 1)(b^{(p-1)/2} + 1),$$

we must have $b^{(p-1)/2} \equiv -1 \bmod p$. □

Corollary 29 *If p is an odd prime, then -1 is a quadratic residue of p if $p \equiv 1 \bmod 4$ and a quadratic nonresidue of p if $p \equiv 3 \bmod 4$.*

Euler's criterion may also be used to determine for what primes 2 is a quadratic residue:

Proposition 30 *For any odd prime p, 2 is a quadratic residue of p if $p \equiv \pm 1 \bmod 8$ and a quadratic nonresidue if $p \equiv \pm 3 \bmod 8$.*

Proof Let A denote the set of all even integers a such that $p/2 < a < p$, and let B denote the set of all even integers b such that $0 < b < p/2$. Since $A \cup B$ is the set of all positive even integers less than p, it has cardinality $r := (p-1)/2$. Evidently $a \in A$ if and only if $p - a$ is odd and $0 < p - a < p/2$. Hence the integers $1, 2, \ldots, r$ are just the elements of B, together with the integers $p - a (a \in A)$. If we denote the cardinality of A by $\#A$, it follows that

$$\begin{aligned} r! &= \prod_{a \in A} (p - a) \prod_{b \in B} b \\ &\equiv (-1)^{\#A} \prod_{a \in A} a \prod_{b \in B} b \bmod p \\ &= (-1)^{\#A} 2^r r! \end{aligned}$$

Thus $2^r \equiv (-1)^{\#A} \bmod p$ and hence, by Proposition 28, 2 is a quadratic residue or nonresidue of p according as $\#A$ is even or odd. But $\#A = k$ if $p = 4k + 1$ and $\#A = k + 1$ if $p = 4k + 3$. The result follows. □

We now introduce some simple group-theoretical concepts. Let G be a finite group and $a \in G$. Then there exist $j, k \in \mathbb{N}$ with $j < k$ such that $a^j = a^k$. Thus $a^{k-j} = 1$, where 1 is the identity element of G. The *order* of a is the least positive integer d such that $a^d = 1$.

Lemma 31 *Let G be a finite group of order n and a an element of G of order d. Then*

(i) *for any $k \in \mathbb{N}$, $a^k = 1$ if and only if d divides k;*

(ii) *for any* $k \in \mathbb{N}$, a^k *has order* $d/(k,d)$;
(iii) $H = \{1, a, \ldots, a^{d-1}\}$ *is a subgroup of* G *and* d *divides* n.

Proof Any $k \in \mathbb{N}$ can be written in the form $k = qd + r$, where $q \geq 0$ and $0 \leq r < d$. Since $a^{qd} = (a^d)^q = 1$, we have $a^k = 1$ if and only if $a^r = 1$, i.e. if and only if $r = 0$, by the definition of d.

It follows that if a^k has order e, then $ke = [k,d]$. Since $[k,d] = kd/(k,d)$, this implies $e = d/(k,d)$. In particular, a^k again has order d if and only if $(k,d) = 1$.

If $0 \leq j, k < d$, put $i = j + k$ if $j + k < d$ and $i = j + k - d$ if $j + k \geq d$. Then $a^j a^k = a^i$, and so H contains the product of any two of its elements. If $0 < k < d$, then $a^k a^{d-k} = 1$, and so H contains also the inverse of any one of its elements. Finally d divides n, by Lagrange's theorem that the order of a subgroup divides the order of the whole group. □

The subgroup H in Lemma 31 is the *cyclic subgroup generated by* a. For $G = \mathbb{Z}_{(m)}^{\times}$, the case which we will be interested in, there is no need to appeal to Lagrange's theorem, since $\mathbb{Z}_{(m)}^{\times}$ has order $\varphi(m)$ and d divides $\varphi(m)$, by Proposition 25 and Lemma 31(i).

A group G is *cyclic* if it coincides with the cyclic subgroup generated by one of its elements. For example, the n-th roots of unity in $\mathbb{C}$ form a cyclic group generated by $e^{2\pi i/n}$. In fact the generators of this group are just the primitive n-th roots of unity.

Our next result provides a sufficient condition for a finite group to be cyclic.

Lemma 32 *A finite group* G *of order* n *is cyclic if, for each positive divisor* d *of* n, *there are at most* d *elements of* G *whose order divides* d.

Proof If H is a cyclic subgroup of G, then its order d divides n. Since all its elements are of order dividing d, the hypothesis of the lemma implies that any element of G whose order divides d must be in H. Furthermore, H contains exactly $\varphi(d)$ elements of order d since, if a generates H, a^k has order d if and only if $(k,d) = 1$.

For each divisor d of n, let $\psi(d)$ denote the number of elements of G of order d. Then, by what we have just proved, either $\psi(d) = 0$ or $\psi(d) = \varphi(d)$. But $\sum_{d|n} \psi(d) = n$, since the order of each element is a divisor of n, and $\sum_{d|n} \varphi(d) = n$, by Proposition 24. Hence we must have $\psi(d) = \varphi(d)$ for every $d|n$. In particular, the group G has $\psi(n) = \varphi(n)$ elements of order n. □

The condition of Lemma 32 is also necessary. For let G be a finite cyclic group of order n, generated by the element a, and let d be a divisor of n. An element $x \in G$ has order dividing d if and only if $x^d = 1$. Thus the elements a^k of G of order dividing d are given by $k = jn/d$, with $j = 0, 1, \ldots, d-1$.

We now return from group theory to number theory.

Proposition 33 *For any prime* p, *the multiplicative group* $\mathbb{F}_p^{\times}$ *of the field* $\mathbb{F}_p$ *is cyclic.*

Proof Put $G = \mathbb{F}_p^{\times}$ and denote the order of G by n. For any divisor d of n, the polynomial $t^d - 1$ has at most d roots in $\mathbb{F}_p$. Hence there are at most d elements of G whose order divides d. The result now follows from Lemma 32. □

The same argument shows that, for an arbitrary field K, any finite subgroup of the multiplicative group of K is cyclic.

In the terminology of number theory, an integer which generates $\mathbb{Z}_{(m)}^{\times}$ is said to be a *primitive root* of m. Primitive roots may be used to replace multiplications mod m by additions mod $\varphi(m)$ in the same way that logarithms were once used in analysis. If g is a primitive root of m, then the elements of $\mathbb{Z}_{(m)}^{\times}$ are precisely $1, g, g^2, \ldots, g^{n-1}$, where $n = \varphi(m)$. Thus for each $a \in \mathbb{Z}_{(m)}^{\times}$ we have $a \equiv g^{\alpha} \bmod m$ for a unique index α $(0 \leq \alpha < n)$. We can construct a table of these indices once and for all. If $a \equiv g^{\alpha}$ and $b \equiv g^{\beta}$, then $ab \equiv g^{\alpha+\beta}$. By replacing $\alpha + \beta$ by its least non-negative residue γ mod n and going backwards in our table we can determine c such that $ab \equiv c \bmod m$.

For any prime p, an essentially complete proof for the existence of primitive roots of p was given by Euler (1774). Jacobi (1839) constructed tables of indices for all primes less than 1000.

We now use primitive roots to prove a general property of polynomials with coefficients from a finite field:

Proposition 34 *If $f(x_1, \ldots, x_n)$ is a polynomial of degree less than n in n variables with coefficients from the finite field $\mathbb{F}_p$, then the number of zeros of f in $\mathbb{F}_p^n$ is divisible by the characteristic p. In particular, $(0, \ldots, 0)$ is not the only zero of f if f has no constant term.*

Proof Put $K = \mathbb{F}_p$ and $g = 1 - f^{p-1}$. If $\alpha = (a_1, \ldots, a_n)$ is a zero of f, then $g(\alpha) = 1$. If α is not a zero of f, then $f(\alpha)^{p-1} = 1$ and $g(\alpha) = 0$. Hence the number N of zeros of f satisfies

$$N \equiv \sum_{\alpha \in K^n} g(\alpha) \bmod p.$$

We will complete the proof by showing that

$$\sum_{\alpha \in K^n} g(\alpha) = 0.$$

Since g has degree less than $n(p-1)$, it is a constant linear combination of polynomials of the form $x_1^{k_1} \cdots x_n^{k_n}$, where $k_1 + \cdots + k_n < n(p-1)$. Thus $k_j < p-1$ for at least one j. Since

$$\sum_{\alpha \in K^n} a_1^{k_1} \cdots a_n^{k_n} = \left(\sum_{a_1 \in K} a_1^{k_1} \right) \cdots \left(\sum_{a_n \in K} a_n^{k_n} \right),$$

it is enough to show that $S_k := \sum_{a \in K} a^k$ is zero for $0 \leq k < p-1$. If $k = 0$, then $a^k = 1$ and $S_0 = p \cdot 1 = 0$. Suppose $1 \leq k < p-1$ and let b be a generator for the multiplicative group $K^{\times}$ of K. Then $c := b^k \neq 1$ and

$$S_k = \sum_{j=1}^{p-1} c^j = c(c^{p-1} - 1)/(c-1) = 0. \qquad \square$$

The general case of Proposition 34 was first proved by Warning (1936), after the particular case had been proved by Chevalley (1936). As an illustration, the particular case implies that, for any integers a, b, c and any prime p, the congruence $ax^2 + by^2 + cz^2 \equiv 0 \bmod p$ has a solution in integers x, y, z not all divisible by p.

If m is not a prime, then $\mathbb{Z}_{(m)}$ is not a field. However, we now show that the group $\mathbb{Z}_{(m)}^{\times}$ is cyclic also if $m = p^2$ is the square of a prime.

Let g be a primitive root of p. It follows from the binomial theorem that

$$(g + p)^p \equiv g^p \bmod p^2.$$

Hence, if $g^p \equiv g \bmod p^2$, then $(g + p)^p \not\equiv g + p \bmod p^2$. Thus, by replacing g by $g + p$ if necessary, we may assume that $g^{p-1} \not\equiv 1 \bmod p^2$. If the order of g in $\mathbb{Z}_{(p^2)}^{\times}$ is d, then d divides $\varphi(p^2) = p(p-1)$. But $\varphi(p) = p-1$ divides d, since $g^d \equiv 1 \bmod p^2$ implies $g^d \equiv 1 \bmod p$ and g is a primitive root of p. Since p is prime and $d \neq p-1$, it follows that $d = p(p-1)$, i.e. $\mathbb{Z}_{(p^2)}^{\times}$ is cyclic with g as generator.

We briefly state some further results about primitive roots, although we will not use them. Gauss (*D.A.*, §89–92) showed that *the group* $\mathbb{Z}_{(m)}^{\times}$ *is cyclic if and only if* $m \in \{2, 4, p^k, 2p^k\}$, where p is an odd prime and $k \in \mathbb{N}$. Evidently 1 is a primitive root of 2 and 3 is a primitive root of 4. *If g is a primitive root of p^2, where p is an odd prime, then g is a primitive root of p^k for every $k \in \mathbb{N}$; and if $g' = g$ or $g + p^k$, according as g is odd or even, then g' is a primitive root of $2p^k$.*

By Fermat's little theorem, if p is prime, then $a^{p-1} \equiv 1 \bmod p$ for every $a \in \mathbb{Z}$ such that $(a, p) = 1$. With the aid of primitive roots we will now show that there exist also composite integers n such that $a^{n-1} \equiv 1 \bmod n$ for every $a \in \mathbb{Z}$ such that $(a, n) = 1$.

Proposition 35 *For any integer $n > 1$, the following two statements are equivalent:*

(i) *$a^{n-1} \equiv 1 \bmod n$ for every integer a such that $(a, n) = 1$;*
(ii) *n is a product of distinct primes and, for each prime $p|n$, $p - 1$ divides $n - 1$.*

Proof Suppose first that (i) holds and assume that, for some prime p, $p^2|n$. As we have just proved, there exists a primitive root g of p^2. Evidently $p \nmid g$. It is easily seen that there exists $c \in \mathbb{N}$ such that $a = g + cp^2$ is relatively prime to n; in fact we can take c to be the product of the distinct prime factors of n, other than p, which do not divide g. Since n divides $a^{n-1} - 1$, also p^2 divides $a^{n-1} - 1$. But a, like g, is a primitive root of p^2, and so its order in $\mathbb{Z}_{(p^2)}^{\times}$ is $\varphi(p^2) = p(p-1)$. Hence $p(p-1)$ divides $n - 1$. But this contradicts $p|n$.

Now let p be any prime divisor of n and let g be a primitive root of p. In the same way as before, there exists $c \in \mathbb{N}$ such that $a = g + cp$ is relatively prime to n. Arguing as before, we see that $\varphi(p) = p - 1$ divides $n - 1$. This proves that (i) implies (ii).

Suppose next that (ii) holds and let a be any integer relatively prime to n. If p is a prime factor of n, then $p \nmid a$ and hence $a^{p-1} \equiv 1 \bmod p$. Since $p - 1$ divides $n - 1$, it follows that $a^{n-1} \equiv 1 \bmod p$. Thus $a^{n-1} - 1$ is divisible by each prime factor of n and hence, since n is squarefree, also by n itself. □

Proposition 35 was proved by Carmichael (1910), and a composite integer n with the equivalent properties stated in the proposition is said to be a *Carmichael number*.

Any Carmichael number n must be odd, since it has an odd prime factor p such that $p-1$ divides $n-1$. Furthermore a Carmichael number must have more than two prime factors. For assume $n = pq$, where $1 < p < q < n$ and $q - 1$ divides $n - 1$. Since $q \equiv 1 \bmod(q-1)$, it follows that

$$0 \equiv pq - 1 \equiv p - 1 \bmod(q-1),$$

which contradicts $p < q$.

The composite integer $561 = 3 \times 11 \times 17$ is a Carmichael number, since 560 is divisible by 2,10 and 16, and it is in fact the smallest Carmichael number. The taxicab number 1729, which Hardy reckoned to Ramanujan was uninteresting, is also a Carmichael number, since $1729 = 7 \times 13 \times 19$. Indeed it is not difficult to show that if p, $2p-1$ and $3p-2$ are all primes, with $p > 3$, then their product is a Carmichael number. Recently Alford, Granville and Pomerance (1994) confirmed a long-standing conjecture by proving that there are infinitely many Carmichael numbers.

Our next topic is of greater importance. Many arithmetical problems require for their solution the determination of an integer which is congruent to several given integers according to various given moduli. We consider first a simple, but important, special case.

Proposition 36 *Let $m = m'm''$, where m' and m'' are relatively prime integers. Then, for any integers a', a'', there exists an integer a, which is uniquely determined* mod m, *such that*

$$a \equiv a' \bmod m', \quad a \equiv a'' \bmod m''.$$

Moreover, a is relatively prime to m if and only if a' is relatively prime to m' and a'' is relatively prime to m''.

Proof By Proposition 22, there exist integers c', c'' such that

$$c'm'' \equiv 1 \bmod m', \quad c''m' \equiv 1 \bmod m''.$$

Thus $e' := c'm''$ is congruent to $1 \bmod m'$ and congruent to $0 \bmod m''$. Similarly $e'' := c''m'$ is congruent to $0 \bmod m'$ and congruent to $1 \bmod m''$. It follows that $a = a'e' + a''e''$ is congruent to $a' \bmod m'$ and congruent to $a'' \bmod m''$.

It is evident that if $b \equiv a \bmod m$, then also $b \equiv a' \bmod m'$ and $b \equiv a'' \bmod m''$. Conversely, if b satisfies these two congruences, then $b - a \equiv 0 \bmod m'$ and $b - a \equiv 0 \bmod m''$. Hence $b - a \equiv 0 \bmod m$, by Proposition 3(i).

Since m' and m'' are relatively prime, it follows from Proposition 3(iv) that $(a, m) = 1$ if and only if $(a, m') = (a, m'') = 1$. Since $a \equiv a' \bmod m'$ implies $(a, m') = (a', m')$, and $a \equiv a'' \bmod m''$ implies $(a, m'') = (a'', m'')$, this proves the last statement of the proposition. □

In algebraic terms, Proposition 36 says that if $m = m'm''$, where m' and m'' are relatively prime integers, then the ring $\mathbb{Z}_{(m)}$ is (isomorphic to) the direct sum of the rings $\mathbb{Z}_{(m')}$ and $\mathbb{Z}_{(m'')}$. Furthermore, the group $\mathbb{Z}_{(m)}^{\times}$ is (isomorphic to) the direct product of the groups $\mathbb{Z}_{(m')}^{\times}$ and $\mathbb{Z}_{(m'')}^{\times}$.

Proposition 36 can be considerably generalized:

Proposition 37 *For any integers $m_1, \ldots, m_n$ and $a_1, \ldots, a_n$, the simultaneous congruences*

$$x \equiv a_1 \bmod m_1, \ldots, x \equiv a_n \bmod m_n$$

have a solution x if and only if

$$a_j \equiv a_k \bmod (m_j, m_k) \quad \textit{for } 1 \leq j < k \leq n.$$

Moreover, y is also a solution if and only if

$$y \equiv x \bmod [m_1, \ldots, m_n].$$

proof The necessity of the conditions is trivial. For if x is a solution and if $d_{jk} = (m_j, m_k)$ is the greatest common divisor of m_j and m_k, then $a_j \equiv x \equiv a_k \bmod d_{jk}$. Also, if y is another solution, then $y - x$ is divisible by $m_1, \ldots, m_n$ and hence also by their least common multiple $[m_1, \ldots, m_n]$.

We prove the sufficiency of the conditions by induction on n. Suppose first that $n = 2$ and $a_1 \equiv a_2 \bmod d$, where $d = (m_1, m_2)$. By the Bézout identity,

$$d = x_1 m_1 - x_2 m_2$$

for some $x_1, x_2 \in \mathbb{Z}$. Since $a_1 - a_2 = kd$ for some $k \in \mathbb{Z}$, it follows that

$$x := a_1 - k x_1 m_1 = a_2 - k x_2 m_2$$

is a solution.

Suppose next that $n > 2$ and the result holds for all smaller values of n. Then there exists $x' \in \mathbb{Z}$ such that

$$x' \equiv a_i \bmod m_i \quad \text{for } 1 \leq i < n,$$

and x' is uniquely determined mod m', where $m' = [m_1, \ldots, m_{n-1}]$. Since any solution of the two congruences

$$x \equiv x' \bmod m', x \equiv a_n \bmod m_n$$

is a solution of the given congruences, we need only show that $x' \equiv a_n \bmod (m', m_n)$. But, by the distributive law connecting greatest common divisors and least common multiples,

$$(m', m_n) = [(m_1, m_n), \ldots, (m_{n-1}, m_n)].$$

Since $x' \equiv a_i \equiv a_n \bmod (m_i, m_n)$ for $1 \leq i < n$, it follows that $x' \equiv a_n \bmod (m', m_n)$. □

Corollary 38 *Let $m_1, \ldots, m_n$ be integers, any two of which are relatively prime, and let $m = m_1 \cdots m_n$ be their product. Then, for any given integers $a_1, \ldots, a_n$, there is a unique integer $x \bmod m$ such that*

$$x \equiv a_1 \bmod m_1, \ldots, x \equiv a_n \bmod m_n.$$

Moreover, x is relatively prime to m if and only if a_i is relatively prime to m_i for $1 \leq i \leq n$.

Corollary 38 can also be proved by an extension of the argument used to prove Proposition 36. Both Proposition 37 and Corollary 38 are referred to as the *Chinese remainder theorem*. Sunzi (4th century A.D.) gave a procedure for obtaining the solution $x = 23$ of the simultaneous congruences

$$x \equiv 2 \bmod 3, \quad x \equiv 3 \bmod 5, \quad x \equiv 2 \bmod 7.$$

Qin Jiushao (1247) gave a general procedure for solving simultaneous congruences, the moduli of which need not be pairwise relatively prime, although he did not state the necessary condition for the existence of a solution. The problem appears to have its origin in the construction of calendars.

6 Sums of Squares

Which positive integers n can be represented as a sum of two squares of integers? The question is answered completely by the following proposition, which was stated by Girard (1625). Fermat (1645) claimed to have a proof, but the first published proof was given by Euler (1754).

Proposition 39 *A positive integer n can be represented as a sum of two squares if and only if for each prime* $p \equiv 3 \bmod 4$ *that divides n, the highest power of p dividing n is even.*

Proof We observe first that, since

$$(x^2 + y^2)(u^2 + v^2) = (xu + yv)^2 + (xv - yu)^2,$$

any product of sums of two squares is again a sum of two squares.

Suppose $n = x^2 + y^2$ for some integers x, y and that n is divisible by a prime $p \equiv 3 \bmod 4$. Then $x^2 \equiv -y^2 \bmod p$. But -1 is not a square in the field $\mathbb{F}_p$, by Corollary 29. Consequently we must have $y^2 \equiv x^2 \equiv 0 \bmod p$. Thus p divides both x and y. Hence p^2 divides n and $(n/p)^2 = (x/p)^2 + (y/p)^2$. It follows by induction that the highest power of p which divides n is even.

Thus the condition in the statement of the proposition is necessary. Suppose now that this condition is satisfied. Then $n = qm^2$, where q is square-free and the only possible prime divisors of q are 2 and primes $p \equiv 1 \bmod 4$. Since $m^2 = m^2 + 0^2$ and $2 = 1^2 + 1^2$, it follows from our initial observation that n is a sum of two squares if every prime $p \equiv 1 \bmod 4$ is a sum of two squares. Following Gauss (1832), we will prove this with the aid of complex numbers.

A complex number $\gamma = a + bi$ is said to be a *Gaussian integer* if $a, b \in \mathbb{Z}$. The set of all Gaussian integers will be denoted by $\mathcal{G}$. Evidently $\gamma \in \mathcal{G}$ implies $\bar{\gamma} \in \mathcal{G}$, where $\bar{\gamma} = a - bi$ is the complex conjugate of γ. Moreover $\alpha, \beta \in \mathcal{G}$ implies $\alpha \pm \beta \in \mathcal{G}$ and $\alpha\beta \in \mathcal{G}$. Thus $\mathcal{G}$ is a commutative ring. In fact $\mathcal{G}$ is an integral domain, since it is a subset of the field $\mathbb{C}$. We are going to show that $\mathcal{G}$ can be given the structure of a Euclidean domain.

Define the *norm* of a complex number $\gamma = a + bi$ to be

$$N(\gamma) = \gamma\bar{\gamma} = a^2 + b^2.$$

Then $N(\gamma) \geq 0$, with equality if and only if $\gamma = 0$, and $N(\gamma_1\gamma_2) = N(\gamma_1)N(\gamma_2)$. If $\gamma \in \mathscr{G}$, then $N(\gamma)$ is an ordinary integer. Furthermore, γ is a unit in $\mathscr{G}$, i.e. γ divides 1 in $\mathscr{G}$, if and only if $N(\gamma) = 1$.

We wish to show that if $\alpha, \beta \in \mathscr{G}$ and $\alpha \neq 0$, then there exist $\kappa, \rho \in \mathscr{G}$ such that

$$\beta = \kappa\alpha + \rho, \quad N(\rho) < N(\alpha).$$

We have $\beta\alpha^{-1} = r + si$, where $r, s \in \mathbb{Q}$. Choose $a, b \in \mathbb{Z}$ so that

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

If $\kappa = a + bi$, then $\kappa \in \mathscr{G}$ and

$$N(\beta\alpha^{-1} - \kappa) \leq 1/4 + 1/4 = 1/2 < 1.$$

Hence if $\rho = \beta - \kappa\alpha$, then $\rho \in \mathscr{G}$ and $N(\rho) < N(\alpha)$.

It follows that we can apply to $\mathscr{G}$ the whole theory of divisibility in a Euclidean domain. Now let p be a prime such that $p \equiv 1 \bmod 4$. We will show that p is a sum of two squares by constructing $\beta \in \mathscr{G}$ for which $N(\beta) = p$.

By Corollary 29, there exists an integer a such that $a^2 \equiv -1 \bmod p$. Put $\alpha = a + i$. Then $N(\alpha) = \alpha\bar{\alpha} = a^2 + 1$ is divisible by p in $\mathbb{Z}$ and hence also in $\mathscr{G}$. However, neither α nor $\bar{\alpha}$ is divisible by p in $\mathscr{G}$, since αp^{-1} and $\bar{\alpha} p^{-1}$ are not in $\mathscr{G}$. Thus p is not a prime in $\mathscr{G}$ and consequently, since $\mathscr{G}$ is a Euclidean domain, it has a factorization $p = \beta\gamma$, where neither β nor γ is a unit. Hence $N(\beta) > 1$, $N(\gamma) > 1$. Since

$$N(\beta)N(\gamma) = N(p) = p^2,$$

it follows that $N(\beta) = N(\gamma) = p$. □

Proposition 39 solves the problem of representing a positive integer as a sum of two squares. What if we allow more than two squares? When congruences were first introduced in §5, it was observed that a positive integer $a \equiv 7 \bmod 8$ could not be represented as a sum of three squares. It was first completely proved by Gauss (1801) that a positive integer can be represented as a sum of three squares if and only if it is not of the form $4^n a$, where $n \geq 0$ and $a \equiv 7 \bmod 8$. The proof of this result is more difficult, and will be given in Chapter VII.

It was conjectured by Bachet (1621) that *every* positive integer can be represented as a sum of four squares. Fermat claimed to have a proof, but the first published proof was given by Lagrange (1770), using earlier ideas of Euler (1751). The proof of the four-squares theorem we will give is similar to that just given for the two-squares theorem, with complex numbers replaced by quaternions.

Proposition 40 *Every positive integer n can be represented as a sum of four squares.*

Proof A quaternion $\gamma = a + bi + cj + dk$ will be said to be a *Hurwitz integer* if a, b, c, d are either all integers or all halves of odd integers. The set of all Hurwitz integers will be denoted by $\mathscr{H}$. Evidently $\gamma \in \mathscr{H}$ implies $\bar{\gamma} \in \mathscr{H}$, where $\bar{\gamma} = a - bi - cj - dk$. Moreover $\alpha, \beta \in \mathscr{H}$ implies $\alpha \pm \beta \in \mathscr{H}$. We will show that $\alpha, \beta \in \mathscr{H}$ also implies $\alpha\beta \in \mathscr{H}$.

Evidently $\gamma \in \mathscr{H}$ if and only if it can be written in the form $\gamma = a_0 h + a_1 i + a_2 j + a_3 k$, where $a_0, a_1, a_2, a_3 \in \mathbb{Z}$ and $h = (1 + i + j + k)/2$. It is obvious that the product of h with i, j or k is again in $\mathscr{H}$ and it is easily verified that $h^2 = h - 1$. It follows that $\mathscr{H}$ is closed under multiplication and hence is a ring.

Define the *norm* of a quaternion $\gamma = a + bi + cj + dk$ to be

$$N(\gamma) = \gamma\bar{\gamma} = a^2 + b^2 + c^2 + d^2.$$

Then $N(\gamma) \geq 0$, with equality if and only if $\gamma = 0$. Moreover, since $\overline{\gamma_1\gamma_2} = \bar{\gamma}_2\bar{\gamma}_1$,

$$N(\gamma_1\gamma_2) = \gamma_1\gamma_2\bar{\gamma}_2\bar{\gamma}_1 = \gamma_1\bar{\gamma}_1\gamma_2\bar{\gamma}_2 = N(\gamma_1)N(\gamma_2).$$

If $\gamma \in \mathscr{H}$, then $N(\gamma) = \gamma\bar{\gamma} \in \mathscr{H}$ and hence $N(\gamma)$ is an ordinary integer. Furthermore, γ is a unit in $\mathscr{H}$, i.e. γ divides 1 in $\mathscr{H}$, if and only if $N(\gamma) = 1$.

We now show that a Euclidean algorithm may be defined on $\mathscr{H}$. Suppose $\alpha, \beta \in \mathscr{H}$ and $\alpha \neq 0$. Then

$$\beta\alpha^{-1} = r_0 + r_1 i + r_2 j + r_3 k,$$

where $r_0, r_1, r_2, r_3 \in \mathbb{Q}$. If $\kappa = a_0 h + a_1 i + a_2 j + a_3 k$, then

$$\begin{aligned}\beta\alpha^{-1} - \kappa = (r_0 - a_0/2) + (r_1 - a_0/2 - a_1)i + (r_2 - a_0/2 - a_2)j \\ + (r_3 - a_0/2 - a_3)k.\end{aligned}$$

We can choose $a_0 \in \mathbb{Z}$ so that $|2r_0 - a_0| \leq 1/2$ and then choose $a_\nu \in \mathbb{Z}$ so that $|r_\nu - a_0/2 - a_\nu| \leq 1/2$ $(\nu = 1, 2, 3)$. Then $\kappa \in \mathscr{H}$ and

$$N(\beta\alpha^{-1} - \kappa) \leq 1/16 + 3/4 = 13/16 < 1.$$

Thus if we set $\rho = \beta - \kappa\alpha$, then $\rho \in \mathscr{H}$ and

$$N(\rho) = N(\beta\alpha^{-1} - \kappa)N(\alpha) < N(\alpha).$$

By repeating this division process finitely many times we see that any $\alpha, \beta \in \mathscr{H}$ have a *greatest common right divisor* $\delta = (\alpha, \beta)_r$. Furthermore, there is a *left Bézout identity*: $\delta = \xi\alpha + \eta\beta$ for some $\xi, \eta \in \mathscr{H}$.

If a positive integer n is a sum of four squares, say $n = a^2 + b^2 + c^2 + d^2$, then $n = \gamma\bar{\gamma}$, where $\gamma = a + bi + cj + dk \in \mathscr{H}$. Since the norm of a product is the product of the norms, it follows that any product of sums of four squares is again a sum of four squares. Hence to prove the proposition we need only show that any prime p is a sum of four squares.

We show first that there exist integers a, b such that $a^2 + b^2 \equiv -1 \bmod p$. This follows from the illustration given for Proposition 34, but we will give a direct proof.

If $p = 2$, we can take $a = 1, b = 0$. If $p \equiv 1 \bmod 4$ then, by Corollary 29, there exists an integer a such that $a^2 \equiv -1 \bmod p$ and we can take $b = 0$. Suppose now that $p \equiv 3 \bmod 4$. Let c be the least positive quadratic non-residue of p. Then $c \geq 2$ and $c - 1$ is a quadratic residue of p. On the other hand, -1 is a quadratic non-residue of p, by Corollary 29. Hence, by Proposition 28, $-c$ is a quadratic residue. Thus there exist integers a, b such that

$$a^2 \equiv -c, b^2 \equiv c - 1 \bmod p,$$

and then $a^2 + b^2 \equiv -1 \bmod p$.

Put $\alpha = 1 + ai + bj$. Then p divides $N(\alpha) = \alpha\bar{\alpha} = 1 + a^2 + b^2$ in $\mathbb{Z}$ and hence also in $\mathscr{H}$. However, p does not divide either α or $\bar{\alpha}$ in $\mathscr{H}$, since αp^{-1} and $\bar{\alpha} p^{-1}$ are not in $\mathscr{H}$.

Let $\gamma = (p, \alpha)_r$. Then $p = \beta\gamma$ for some $\beta \in \mathscr{H}$. If β were a unit, p would be a right divisor of γ and hence also of α, which is a contradiction. Therefore $N(\beta) > 1$. Evidently $\gamma\bar{\alpha}$ is a common right divisor of $p\bar{\alpha}$ and $\alpha\bar{\alpha}$, and the Bézout representation for γ implies that $\gamma\bar{\alpha} = (p\bar{\alpha}, \alpha\bar{\alpha})_r$. Since $p\bar{\alpha} = \bar{\alpha}p$ and p divides $\alpha\bar{\alpha}$, it follows that p is a right divisor of $\gamma\bar{\alpha}$. Since p does not divide $\bar{\alpha}$, γ is not a unit and hence $N(\gamma) > 1$. Since

$$N(\beta)N(\gamma) = N(p) = p^2,$$

we must have $N(\beta) = N(\gamma) = p$.

Thus if $\gamma = c_0 + c_1 i + c_2 j + c_3 k$, then $c_0^2 + c_1^2 + c_2^2 + c_3^2 = p$. If $c_0, \ldots, c_3$ are all integers, we are finished. Otherwise $c_0, \ldots, c_3$ are all halves of odd integers. Hence we can write $c_\nu = 2d_\nu + e_\nu$, where $d_\nu \in \mathbb{Z}$ and $e_\nu = \pm 1/2$. If we put

$$\delta = d_0 + d_1 i + d_2 j + d_3 k, \quad \varepsilon = e_0 + e_1 i + e_2 j + e_3 k,$$

then $\gamma = 2\delta + \varepsilon$ and $N(\varepsilon) = 1$. Hence $\theta := \gamma\bar{\varepsilon} = 2\delta\bar{\varepsilon} + 1$ has all its coordinates integers and $N(\theta) = N(\gamma) = p$. □

In his *Meditationes Algebraicae*, which also contains the first statement in print of Wilson's theorem, Waring (1770) stated that every positive integer is a sum of at most 4 positive integral squares, of at most 9 positive integral cubes and of at most 19 positive integral fourth powers. The statement concerning squares was proved by Lagrange in the same year, as we have seen. The statement concerning cubes was first proved by Wieferich (1909), with a gap filled by Kempner (1912), and the statement concerning fourth powers was first proved by Balasubramanian, Deshouillers and Dress (1986).

In a later edition of his book, Waring (1782) raised the same question for higher powers. *Waring's problem* was first solved by Hilbert (1909), who showed that, for each $k \in \mathbb{N}$, there exists $\gamma_k \in \mathbb{N}$ such that every positive integer is a sum of at most γ_k k-th powers. The least possible value of γ_k is traditionally denoted by $g(k)$. For example, $g(2) = 4$, since $7 = 2^2 + 3 \cdot 1^2$ is not a sum of less than 4 squares.

A lower bound for $g(k)$ was already derived by Euler (c. 1772). Let $m = \lfloor (3/2)^k \rfloor$ denote the greatest integer $\leq (3/2)^k$ and take

$$n = 2^k m - 1.$$

Since $1 \leq n < 3^k$, the only k-th powers of which n can be the sum are 0^k, 1^k and 2^k. Since the number of powers 2^k must be less than m, and since $n = (m-1)2^k + (2^k - 1)1^k$, the least number of k-th powers with sum n is $m + 2^k - 2$. Hence $g(k) \geq w(k)$, where

$$w(k) = \lfloor (3/2)^k \rfloor + 2^k - 2.$$

In particular,

$$w(2) = 4, \quad w(3) = 9, \quad w(4) = 19, \quad w(5) = 37, \quad w(6) = 73.$$

By the results stated above, $g(k) = w(k)$ for $k = 2, 3, 4$ and this has been shown to hold also for $k = 5$ by Chen (1964) and for $k = 6$ by Pillai (1940).

Hilbert's method of proof yielded rather large upper bounds for $g(k)$. A completely new approach was developed in the 1920's by Hardy and Littlewood, using their analytic 'circle' method. They showed that, for each $k \in \mathbb{N}$, there exists $\Gamma_k \in \mathbb{N}$ such that every sufficiently large positive integer is a sum of at most Γ_k k-th powers. The least possible value of Γ_k is traditionally denoted by $G(k)$. For example, $G(2) = 4$, since no positive integer $n \equiv 7 \bmod 8$ is a sum of less than four squares. Davenport (1939) showed that $G(4) = 16$, but these are the only two values of k for which today $G(k)$ is known exactly.

It is obvious that $G(k) \leq g(k)$, and in fact $G(k) < g(k)$ for all $k > 2$. In particular, Dickson (1939) showed that 23 and 239 are the only positive integers which require the maximum 9 cubes. Hardy and Littlewood obtained the upper bound $G(k) \leq (k-2)2^{k-1}+5$, but this has been repeatedly improved by Hardy and Littlewood themselves, Vinogradov and others. For example, Wooley (1992) has shown that $G(k) \leq k(\log k + \log\log k + O(1))$.

By using the upper bound for $G(k)$ of Vinogradov (1935), it was shown by Dickson, Pillai and Niven (1936–1944) that $g(k) = w(k)$ for any given $k > 6$, *provided that*

$$(3/2)^k - \lfloor (3/2)^k \rfloor \leq 1 - \lfloor (3/2)^k \rfloor / 2^k.$$

It is possible that this inequality holds for every $k \in \mathbb{N}$. For a given k, it may be checked by direct calculation, and Kubina and Wunderlich (1990) have verified in this way that the inequality holds if $k \leq 471600000$. Furthermore, using a p-adic extension by Ridout (1957) of the theorem of Roth (1955) on the approximation of algebraic numbers by rationals, Mahler (1957) proved that there exists $k_0 \in \mathbb{N}$ such that the inequality holds for all $k \geq k_0$. However, the proof does not provide a means of estimating k_0.

Thus we have the bizarre situation that $G(k)$ is known for only two values of k, that $g(k)$ is known for a vast number of values of k and is given by a simple formula, probably for all k, but the information about $g(k)$ is at present derived from information about $G(k)$. Is it too much to hope that an examination of the numerical data will reveal some pattern in the fractional parts of $(3/2)^k$?

7 Further Remarks

There are many good introductory books on the theory of numbers, e.g. Davenport [4], LeVeque [28] and Scholz [41]. More extensive accounts are given in Hardy and Wright [15], Hua [18], Narkiewicz [33] and Niven *et al.* [34].

Historical information is provided by Dickson [5], Smith [42] and Weil [46], as well as the classics Euclid [11], Gauss [13] and Dirichlet [6]. Gauss's masterpiece is quoted here and in the text as '*D.A.*'

The reader is warned that, besides its use in §1, the word 'lattice' also has quite a different mathematical meaning, which will be encountered in Chapter VIII.

The basic theory of divisibility is discussed more thoroughly than in the usual texts by Stieltjes [43]. For Proposition 6, see Prüfer [35]. In the theory of groups, Schreier's

refinement theorem and the Jordan–Hölder theorem may be viewed as generalizations of Propositions 6 and 7. These theorems are stated and proved in Chapter I, §3 of Lang [23]. The fundamental theorem of arithmetic (Proposition 7) is usually attributed to Gauss (*D.A.*, §16). However, it is really contained in Euclid's *Elements* (Book VII, Proposition 31 and Book IX, Proposition 14), except for the appropriate terminology. Perhaps this is why Euler and his contemporaries simply assumed it without proof.

Generalizations of the fundamental theorem of arithmetic to other algebraic structures are discussed in Chap. 2 of Jacobson [21]. For factorial domains, see Samuel [39].

Our discussion of the fundamental theorem did not deal with the practical problems of deciding if a given integer is prime or composite and, in the latter case, of obtaining its factorization into primes. Evidently if the integer a is composite, its least prime factor p satisfies $p^2 \leq a$. In former days one used this observation in conjunction with tables, such as [24], [25], [26]. With new methods and supercomputers, the primality of integers with hundreds of digits can now be determined without difficulty. The progress in this area may be traced through the survey articles [48], [7] and [27]. Factorization remains a more difficult problem, and this difficulty has found an important application in *public-key cryptography*; see Rivest *et al.* [37].

For Proposition 12, cf. Hillman and Hoggatt [17]. A proof that the ring of all algebraic integers is a Bézout domain is given on p. 86 of Mann [31]. The ring of all functions which are holomorphic in a given region was shown to be a Bézout domain by Wedderburn (1915); see Narasimhan [32].

For Gauss's version of Proposition 17, see *D.A.*, §42. It is natural to ask if Corollary 18 remains valid if the polynomial ring $R[t]$ is replaced by the ring $R[[t]]$ of formal power series. The ring $K[[t_1, \ldots, t_m]]$ of all formal power series in finitely many indeterminates with coefficients from an arbitrary field K is indeed a factorial domain. However, if R is a factorial domain, the integral domain $R[[t]]$ of all formal power series in t with coefficients from R need not be factorial. For an example in which R is actually a complete local ring, see Salmon [38].

For generalizations of Eisenstein's irreducibility criterion (Proposition 19), see Gao [12]. Proposition 21 is proved in Rhai [36]. Euclidean domains are studied further in Samuel [40]. Quadratic fields $\mathbb{Q}(\sqrt{d})$ whose ring of integers $\mathcal{O}_d$ is Euclidean are discussed in Clark [3], Dubois and Steger [8] and Eggleton *et al.* [9].

Congruences are discussed in all the books on number theory cited above. In connection with Lemma 32 we mention a result of Frobenius (1895). Frobenius proved that if G is a finite group of order n and if d is a positive divisor of n, then the number of elements of G whose order divides d is a multiple of d. He conjectured that if the number is exactly d, then these elements form a (normal) subgroup of G. The conjecture can be reduced to the case where G is simple, since a counterexample of minimal order must be a noncyclic simple group. By appealing to the recent classification of all finite simple groups (see Chapter V, §7), the proof of the conjecture was completed by Iiyori and Yamaki [20].

There is a table of primitive roots on pp. 52–56 of Hua [18]. For more extensive tables, see Western and Miller [47].

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 $p = 2, 3$. Artin (1927) conjectured that if the integer a is not

a square or -1, then it is a primitive root for infinitely many primes p. (A quantitative form of the conjecture is considered in Chapter IX.) If the conjecture is not true, then it is almost true, since it has been shown by Heath-Brown [16] that there are at most 3 square-free positive integers a for which it fails.

A finite subgroup of the multiplicative group of a division ring need not be cyclic. For example, if $\mathbb{H}$ is the division ring of Hamilton's quaternions, $\mathbb{H}^\times$ contains the non-cyclic subgroup $\{\pm 1, \pm i, \pm j, \pm k\}$ of order 8. All possible finite subgroups of the multiplicative group of a division ring have been determined (with the aid of *class field theory*) by Amitsur [2].

For Carmichael numbers, see Alford *et al.* [1].

Galois (1830) showed that there were other finite fields besides $\mathbb{F}_p$ and indeed, as Moore (1893) later proved, he found them all. Finite fields have the following basic properties:

(i) The number of elements in a finite field is a prime power p^n, where $n \in \mathbb{N}$ and the prime p is the characteristic of the field.
(ii) For any prime power $q = p^n$, there is a finite field $\mathbb{F}_q$ containing exactly q elements. Moreover the field $\mathbb{F}_q$ is unique, up to isomorphism, and is the splitting field of the polynomial $t^q - t$ over $\mathbb{F}_p$.
(iii) For any finite field $\mathbb{F}_q$, the multiplicative group $\mathbb{F}_q^\times$ of nonzero elements is cyclic.
(iv) If $q = p^n$, the map $\sigma : a \to a^p$ is an automorphism of $\mathbb{F}_q$ and the distinct automorphisms of $\mathbb{F}_q$ are the powers $\sigma^k (k = 0, 1, \ldots, n-1)$.

The theorem of Chevalley and Warning (Proposition 34) extends immediately to arbitrary finite fields. Proofs and more detailed information on finite fields may be found in Lidl and Niederreiter [30] and in Joly [22].

A celebrated theorem of Wedderburn (1905) states that any finite division ring is a field, i.e. the commutative law of multiplication is a consequence of the other field axioms if the number of elements is finite. Here is a purely algebraic proof.

Assume there exists a finite division ring which is not a field and let D be one of minimum cardinality. Let C be the centre of D and $a \in D \backslash C$. The set M of all elements of D which commute with a is a field, since it is a division ring but not the whole of D. Evidently M is a maximal subfield of D which contains a. If $[D : C] = n$ and $[M : C] = m$ then, by Proposition I.32, $[D : M] = m$ and $n = m^2$. Thus m is independent of a.

If C has cardinality q, then D has cardinality q^n, M has cardinality q^m and the number of conjugates of a in D is $(q^n - 1)/(q^m - 1)$. Since this holds for every $a \in D \backslash C$, the partition of the multiplicative group of D into conjugacy classes shows that

$$q^n - 1 = q - 1 + r(q^n - 1)/(q^m - 1)$$

for some positive integer r. Hence $q - 1$ is divisible by

$$(q^n - 1)/(q^m - 1) = 1 + q^m + \cdots + q^{m(m-1)}.$$

Since $n > m > 1$, this is a contradiction.

For the history of the Chinese remainder theorem (not only in China), see Libbrecht [29].

We have developed the arithmetic of quaternions only as far as is needed to prove the four-squares theorem. A fuller account was given in the original (1896) paper of Hurwitz [19]. For more information about sums of squares, see Grosswald [14] and also Chapter XIII. For Waring's problem, see Waring [45], Ellison [10] and Vaughan [44].

8 Selected References

[1] W.R. Alford, A. Granville and C. Pomerance, There are infinitely many Carmichael numbers, *Ann. Math.* **139** (1994), 703–722.

[2] S.A. Amitsur, Finite subgroups of division rings, *Trans. Amer. Math. Soc.* **80** (1955), 361–386.

[3] D.A. Clark, A quadratic field which is Euclidean but not norm-Euclidean, *Manuscripta Math.* **83** (1994), 327–330.

[4] H. Davenport, *The higher arithmetic*, 7th ed., Cambridge University Press, 1999.

[5] L.E. Dickson, *History of the theory of numbers*, 3 vols., Carnegie Institute, Washington, D.C., 1919–1923. [Reprinted, Chelsea, New York, 1992.]

[6] P.G.L. Dirichlet, *Lectures on number theory,* with supplements by R. Dedekind, English transl. by J. Stillwell, American Mathematical Society, Providence, R.I., 1999. [German original, 1894.]

[7] J.D. Dixon, Factorization and primality tests, *Amer. Math. Monthly* **91** (1984), 333–352.

[8] D.W. Dubois and A. Steger, A note on division algorithms in imaginary quadratic fields, *Canad. J. Math.* **10** (1958), 285–286.

[9] R.B. Eggleton, C.B. Lacampagne and J.L. Selfridge, Euclidean quadratic fields, *Amer. Math. Monthly* **99** (1992), 829–837.

[10] W.J. Ellison, Waring's problem, *Amer. Math. Monthly* **78** (1971), 10–36.

[11] Euclid, *The thirteen books of Euclid's elements*, English translation by T.L. Heath, 2nd ed., reprinted in 3 vols., Dover, New York, 1956.

[12] S. Gao, Absolute irreducibility of polynomials via Newton polytopes, *J. Algebra* **237** (2001), 501–520.

[13] C.F. Gauss, *Disquisitiones arithmeticae*, English translation by A.A. Clarke, revised by W.C. Waterhouse, Springer, New York, 1986. [Latin original, 1801.]

[14] E. Grosswald, *Representations of integers as sums of squares*, Springer-Verlag, New York, 1985.

[15] G.H. Hardy and E.M. Wright, *An introduction to the theory of numbers*, 6th ed., Oxford University Press, 2008.

[16] D.R. Heath-Brown, Artin's conjecture for primitive roots, *Quart. J. Math. Oxford Ser.* (2) **37** (1986), 27–38.

[17] A.P. Hillman and V.E. Hoggatt, Exponents of primes in generalized binomial coefficients, *J. Reine Angew. Math.* **262/3** (1973), 375–380.

[18] L.K. Hua, *Introduction to number theory*, English translation by P. Shiu, Springer-Verlag, Berlin, 1982.

[19] A. Hurwitz, Über die Zahlentheorie der Quaternionen, *Mathematische Werke, Band II*, pp. 303–330, Birkhäuser, Basel, 1933.

[20] N. Iiyori and H. Yamaki, On a conjecture of Frobenius, *Bull. Amer. Math. Soc. (N.S)* **25** (1991), 413–416.

[21] N. Jacobson, *Basic Algebra I*, 2nd ed., W.H. Freeman, New York, 1985.

[22] J.-R. Joly, Équations et variétés algébriques sur un corps fini, *Enseign. Math.* (2) **19** (1973), 1–117.

[23] S. Lang, *Algebra*, corrected reprint of 3rd ed., Addison-Wesley, Reading, Mass., 1994.
[24] D.H. Lehmer, *Guide to tables in the theory of numbers*, National Academy of Sciences, Washington, D.C., reprinted 1961.
[25] D.N. Lehmer, *List of prime numbers from* 1 *to* 10,006,721, reprinted, Hafner, New York, 1956.
[26] D.N. Lehmer, *Factor table for the first ten millions*, reprinted, Hafner, New York, 1956.
[27] A.K. Lenstra, Primality testing, *Proc. Symp. Appl. Math.* **42** (1990), 13–25.
[28] W.J. LeVeque, *Fundamentals of number theory*, reprinted Dover, Mineola, N.Y., 1996.
[29] U. Libbrecht, *Chinese mathematics in the thirteenth century*, MIT Press, Cambridge, Mass., 1973.
[30] R. Lidl and H. Niederreiter, *Finite fields*, 2nd ed., Cambridge University Press, 1997.
[31] H.B. Mann, *Introduction to algebraic number theory*, Ohio State University, Columbus, Ohio, 1955.
[32] R. Narasimhan, *Complex analysis in one variable*, Birkhäuser, Boston, Mass., 1985.
[33] W. Narkiewicz, *Number theory*, English translation by S. Kanemitsu, World Scientific, Singapore, 1983.
[34] I. Niven, H.S. Zuckerman and H.L. Montgomery, *An introduction to the theory of numbers*, 5th ed., Wiley, New York, 1991.
[35] H. Prüfer, Untersuchungen über Teilbarkeitseigenschaften, *J. Reine Angew. Math.* **168** (1932), 1–36.
[36] T.-S. Rhai, A characterization of polynomial domains over a field, *Amer. Math. Monthly* **69** (1962), 984–986.
[37] R.L. Rivest, A. Shamir and L. Adleman, A method for obtaining digital signatures and public-key cryptosystems, *Comm. ACM* **21** (1978), 120–126.
[38] P. Salmon, Sulla fattorialità delle algebre graduate e degli anelli locali, *Rend. Sem. Mat. Univ. Padova* **41** (1968), 119–138.
[39] P. Samuel, Unique factorization, *Amer. Math. Monthly* **75** (1968), 945–952.
[40] P. Samuel, About Euclidean rings, *J. Algebra* **19** (1971), 282–301.
[41] A. Scholz, *Einführung in die Zahlentheorie*, revised and edited by B. Schoeneberg, 5th ed., de Gruyter, Berlin, 1973.
[42] H.J.S. Smith, Report on the theory of numbers, *Collected mathematical papers, Vol.1*, pp. 38–364, reprinted, Chelsea, New York, 1965. [Original, 1859–1865.]
[43] T.J. Stieltjes, Sur la théorie des nombres, *Ann. Fac. Sci. Toulouse* **4** (1890), 1–103. [Reprinted in Tome 2, pp. 265–377 of T.J. Stieltjes, *Oeuvres complètes*, 2 vols., Noordhoff, Groningen, 1914–1918.]
[44] R.C. Vaughan, *The Hardy–Littlewood method*, 2nd ed., Cambridge Tracts in Mathematics **125**, Cambridge University Press, 1997.
[45] E. Waring, *Meditationes algebraicae*, English transl. of 1782 edition by D. Weeks, Amer. Math. Soc., Providence, R.I., 1991.
[46] A. Weil, *Number theory*: *an approach through history*, Birkhäuser, Boston, Mass., 1984.
[47] A.E. Western and J.C.P. Miller, *Tables of indices and primitive roots*, Royal Soc. Math. Tables, Vol. 9, Cambridge University Press, London, 1968.
[48] H.C. Williams, Primality testing on a computer, *Ars Combin.* **5** (1978), 127–185.

Additional References

M. Agarwal, N. Kayal and N. Saxena, PRIMES is in P, *Ann. of Math.* **160** (2004), 781–793. [An unconditional deterministic polynomial-time algorithm for determining if an integer > 1 is prime or composite.]
A. Granville, It is easy to determine whether a given integer is prime, *Bull. Amer. Math. Soc. (N.S.)* **42** (2005), 3–38.

III

More on Divisibility

In this chapter the theory of divisibility is developed further. The various sections of the chapter are to a large extent independent. We consider in turn the law of quadratic reciprocity, quadratic fields, multiplicative functions, and linear Diophantine equations.

1 The Law of Quadratic Reciprocity

Let p be an odd prime. An integer a which is not divisible by p is said to be a *quadratic residue*, or *quadratic nonresidue*, of p according as the congruence

$$x^2 \equiv a \bmod p$$

has, or has not, a solution x. We will speak of the *quadratic nature* of $a \bmod p$, meaning whether a is a quadratic residue or nonresidue of p.

Let q be an odd prime different from p. The *law of quadratic reciprocity* connects the quadratic nature of $q \bmod p$ with the quadratic nature of $p \bmod q$. It states that if either p or q is congruent to $1 \bmod 4$, then the quadratic nature of $q \bmod p$ is the same as the quadratic nature of $p \bmod q$, but if both p and q are congruent to $3 \bmod 4$ then the quadratic nature of $q \bmod p$ is different from the quadratic nature of $p \bmod q$.

This remarkable result plays a key role in the arithmetic theory of quadratic forms. It was discovered empirically by Euler (1783). Legendre (1785) gave a partial proof and later (1798) introduced the convenient 'Legendre symbol'. The first complete proofs were given by Gauss (1801) in his *Disquisitiones Arithmeticae*. Indeed the result so fascinated Gauss that during the course of his lifetime he gave eight proofs, four of them resting on completely different principles: an induction argument, the theory of binary quadratic forms, properties of sums of roots of unity, and a combinatorial lemma. The proof we are now going to give is also of a combinatorial nature. Its idea originated with Zolotareff (1872), but our treatment is based on Rousseau (1994).

Let n be a positive integer and let X be the set $\{0, 1, \ldots, n-1\}$. As in §7 of Chapter I, a permutation α of X is said to be *even* or *odd* according as the total number of inversions of order it induces is even or odd. If a is an integer relatively prime to n, then the map $\pi_a \colon X \to X$ defined by

$$\pi_a(x) = ax \bmod n$$

W.A. Coppel, *Number Theory: An Introduction to Mathematics*, Universitext,
DOI: 10.1007/978-0-387-89486-7_3, © Springer Science + Business Media, LLC 2009

is a permutation of X. We define the *Jacobi symbol* (a/n) to be $\operatorname{sgn}(\pi_a)$, i.e.

$$(a/n) = 1 \text{ or } -1$$

according as the permutation π_a is even or odd. Thus $(a/1) = 1$, for every integer a. (The definition is sometimes extended by putting $(a/n) = 0$ if a and n are not relatively prime.)

Proposition 1 *For any positive integer n and any integers a,b relatively prime to n, the Jacobi symbol has the following properties:*

(i) $(1/n) = 1$,
(ii) $(a/n) = (b/n)$ *if* $a \equiv b \bmod n$,
(iii) $(ab/n) = (a/n)(b/n)$,
(iv) $(-1/n) = 1$ *if* $n \equiv 1$ *or* $2 \bmod 4$ *and* $= -1$ *if* $n \equiv 3$ *or* $0 \bmod 4$.

Proof The first two properties follow at once from the definition of the Jacobi symbol. If a and b are both relatively prime to n, then so also is their product ab. Since $\pi_{ab} = \pi_a \pi_b$, we have $\operatorname{sgn}(\pi_{ab}) = \operatorname{sgn}(\pi_a)\operatorname{sgn}(\pi_b)$, which implies (iii). We now evaluate $(-1/n)$. Since the map $\pi_{-1}: x \to -x \bmod n$ fixes 0 and reverses the order of $1, \ldots, n-1$, the total number of inversions of order is $(n-2) + (n-3) + \cdots + 1 = (n-1)(n-2)/2$. It follows that $(-1/n) = (-1)^{(n-1)/2}$ or $(-1)^{(n-2)/2}$ according as n is odd or even. This proves (iv). □

Proposition 2 *For any relatively prime positive integers m, n,*

(i) *if m and n are both odd, then* $(m/n)(n/m) = (-1)^{(m-1)(n-1)/4}$;
(ii) *if m is odd and n even, then* $(m/n) = 1$ *or* $(-1)^{(m-1)/2}$ *according as* $n \equiv 2$ *or* $0 \bmod 4$.

Proof The cyclic permutation $\tau: x \to x + 1 \bmod n$ of the set $X = \{0, 1, \ldots, n-1\}$ has sign $(-1)^{n-1}$, since the number of inversions of order is $n - 1$. Hence, for any integer $b \geq 0$ and any integer a relatively prime to n, the linear permutation

$$\tau^b \pi_a: x \to ax + b \bmod n$$

of X has $\operatorname{sign}(-1)^{b(n-1)}(a/n)$.

Put $Y = \{0, 1, \ldots, m-1\}$ and $P = X \times Y$. We consider two transformations μ and ν of P, defined by

$$\mu(x, y) = (mx + y \bmod n, y), \quad \nu(x, y) = (x, x + ny \bmod m).$$

For each fixed y, μ defines a permutation of the set (X, y) with sign $(-1)^{y(n-1)}(m/n)$. Since $\sum_{y=0}^{m-1} y = m(m-1)/2$, it follows that the permutation μ of P has sign

$$\operatorname{sgn}(\mu) = (-1)^{m(m-1)(n-1)/2}(m/n)^m.$$

Similarly the permutation ν of P has sign

$$\operatorname{sgn}(\nu) = (-1)^{n(m-1)(n-1)/2}(n/m)^n,$$

and hence $\alpha := \nu\mu^{-1}$ has sign

$$\operatorname{sgn}(\alpha) = (-1)^{(m+n)(m-1)(n-1)/2}(m/n)^m(n/m)^n.$$

But α is the permutation $(mx + y \bmod n, y) \to (x, x + ny \bmod m)$ and its sign can be determined directly in the following way.

Put $Z = \{0, 1, \ldots, mn - 1\}$. By Proposition II.36, for any $(x, y) \in P$ there is a unique $z \in Z$ such that

$$z \equiv x \bmod n, \quad z \equiv y \bmod m.$$

Moreover, any $z \in Z$ is obtained in this way from a unique $(x, y) \in P$. For any $z \in Z$, we will denote by $\rho(z)$ the corresponding element of P. Then the permutation α can be written in the form $\rho(mx + y) \to \rho(x + ny)$. Since ρ is a bijective map, the sign of the permutation α of P will be the same as the sign of the permutation $\beta = \rho^{-1}\alpha\rho\colon mx + y \to x + ny$ of Z. An inversion of order for β occurs when both $mx + y > mx' + y'$ and $x + ny < x' + ny'$, i.e. when both $m(x - x') > y' - y$ and $x - x' < n(y' - y)$. But these inequalities imply $mn(x - x') > x - x'$ and hence $x > x'$, $y' > y$. Conversely, if $x > x'$, $y' > y$, then

$$m(x - x') \geq m > y' - y, \quad n(y' - y) \geq n > x - x'.$$

Since the number of $(x, y), (x', y') \in P$ with $x > x'$, $y < y'$ is $m(m-1)/2 \cdot n(n-1)/2$, it follows that the sign of the permutation α is $(-1)^{mn(m-1)(n-1)/4}$. Comparing this expression with the expression previously found, we obtain

$$(m/n)^m(n/m)^n = (-1)^{(mn+2m+2n)(m-1)(n-1)/4}.$$

This simplifies to the first statement of the proposition if m and n are both odd, and to the second statement if m is odd and n even. □

Corollary 3 *For any odd positive integer n,* $(2/n) = 1$ *or* -1 *according as* $n \equiv \pm 1$ *or* $\pm 5 \bmod 8$.

Proof Since the result is already known for $n = 1$, we suppose $n > 1$. Then either n or $n - 2$ is congruent to $1 \bmod 4$ and so, by Proposition 1 and Proposition 2(i),

$$(2/n) = (-1/n)((n-2)/n) = (-1/n)(n/(n-2)) = (-1)^{(n-1)/2}(2/(n-2)).$$

Iterating, we obtain $(2/n) = (-1)^h$, where $h = (n-1)/2 + (n-3)/2 + \cdots + 1 = (n^2 - 1)/8$. The result follows. □

The value of (a/n) when n is even is completely determined by Propositions 1 and 2. The evaluation of (a/n) when n is odd reduces by these propositions and Corollary 3 to the evaluation of (m/n) for odd $m > 1$. Although Proposition 2 does not provide a formula for the Jacobi symbol in this case, it does provide a method for its rapid evaluation, as we now show.

If m and n are relatively prime odd positive integers, we can write $m = 2hn + \varepsilon_1 n_1$, where $h \in \mathbb{Z}$, $\varepsilon_1 = \pm 1$ and n_1 is an odd positive integer less than n. Then n and n_1 are also relatively prime and

$$(m/n) = (\varepsilon_1/n)(n_1/n).$$

If $n_1 = 1$, we are finished. Otherwise, using Proposition 2(i), we obtain

$$(m/n) = (-1)^{(n_1-1)(n-1)/4}(\varepsilon_1/n)(n/n_1) = \pm(n/n_1),$$

where the minus sign holds if and only if n and $\varepsilon_1 n_1$ are both congruent to 3 mod 4. The process can now be repeated with m, n replaced by n, n_1. After finitely many steps the process must terminate with $n_s = 1$.

As an example,

$$\begin{aligned}\left(\frac{2985}{1951}\right) &= \left(\frac{-1}{1951}\right)\left(\frac{917}{1951}\right) = -\left(\frac{1951}{917}\right)\\ &= -\left(\frac{117}{917}\right) = -\left(\frac{917}{117}\right)\\ &= -\left(\frac{-1}{117}\right)\left(\frac{19}{117}\right) = -\left(\frac{117}{19}\right)\\ &= -\left(\frac{3}{19}\right) = \left(\frac{19}{3}\right) = \left(\frac{1}{3}\right) = 1.\end{aligned}$$

Further properties of the Jacobi symbol can be derived from those already established.

Proposition 4 *If n, n' are positive integers and if a is an integer relatively prime to n such that $n' \equiv n \bmod 4a$, then $(a/n') = (a/n)$.*

Proof If $a = -1$ then, since $n' \equiv n \bmod 4$, $(a/n') = (a/n)$, by Proposition 1(iv). If $a = 2$ then, since n and n' are odd and $n' \equiv n \bmod 8$, $(a/n') = (a/n)$, by Corollary 3. Consequently, by Proposition 1(iii), it is sufficient to prove the result for odd $a > 1$.

If n is even, the result now follows from Proposition 2(ii). If n is odd, it follows from Proposition 2(i) and Proposition 1. □

Proposition 5 *If the integer a is relatively prime to the odd positive integers n and n', then $(a/nn') = (a/n)(a/n')$.*

Proof We have $a \equiv a' \bmod nn'$ for some $a' \in \{1, 2, \ldots, nn'\}$. Since nn' is odd, we can choose $j \in \{0, 1, 2, 3\}$ so that $a'' = a' + jnn'$ satisfies $a'' \equiv 1 \bmod 4$. Then, by Propositions 1 and 2,

$$(a/nn') = (a''/nn') = (nn'/a'') = (n/a'')(n'/a'') = (a''/n)(a''/n') = (a/n)(a/n').$$

□

Proposition 5 reduces the evaluation of (a/n) for odd positive n to the evaluation of (a/p), where p is an odd prime. This is where we make the connection with quadratic residues:

Proposition 6 *If p is an odd prime and a an integer not divisible by p, then $(a/p) = 1$ or -1 according as a is a quadratic residue or nonresidue of p. Moreover, exactly half of the integers $1, \ldots, p-1$ are quadratic residues of p.*

Proof If a is a quadratic residue of p, there exists an integer x such that $x^2 \equiv a \bmod p$ and hence

$$(a/p) = (x^2/p) = (x/p)(x/p) = 1.$$

Let g be a primitive root mod p. Then the integers $1, g, \ldots, g^{p-2}$ mod p are just a rearrangement of the integers $1, 2, \ldots, p-1$. The permutation

$$\pi_g : x \to gx \pmod{p}$$

fixes 0 and cyclically permutes the remaining elements $1, g, \ldots, g^{p-2}$. Since the number of inversions of order is $p-2$, it follows that $(g/p) = -1$. For any integer a not divisible by p there is a unique $k \in \{0, 1, \ldots, p-2\}$ such that $a \equiv g^k \bmod p$. Hence

$$(a/p) = (g^k/p) = (g/p)^k = (-1)^k.$$

Thus $(a/p) = 1$ if and only if k is even and then $a \equiv x^2 \bmod p$ with $x = g^{k/2}$.

This proves the first statement of the proposition. Since exactly half the integers in the set $\{0, 1, \ldots, p-2\}$ are even, it also proves again (cf. Proposition II.28) the second statement. □

The law of quadratic reciprocity can now be established without difficulty:

Theorem 7 *Let p and q be distinct odd primes. Then the quadratic natures of $p \bmod q$ and $q \bmod p$ are the same if $p \equiv 1$ or $q \equiv 1 \bmod 4$, but different if $p \equiv q \equiv 3 \bmod 4$.*

Proof The result follows at once from Proposition 6 since, by Proposition 2(i), if either $p \equiv 1$ or $q \equiv 1 \bmod 4$ then $(p/q) = (q/p)$, but if $p \equiv q \equiv 3 \bmod 4$ then $(p/q) = -(q/p)$. □

Legendre (1798) *defined* $(a/p) = 1$ or -1 according as a was a quadratic residue or nonresidue of p, and Jacobi (1837) extended this definition to (a/n) for any odd positive integer n relatively prime to a by setting

$$(a/n) = \prod_p (a/p),$$

where p runs through the prime divisors of n, each occurring as often as its multiplicity. Propositions 5 and 6 show that these definitions of Legendre and Jacobi are equivalent to the definition adopted here. The relations $(-1/p) = (-1)^{(p-1)/2}$ and $(2/p) = (-1)^{(p^2-1)/8}$ for odd primes p are often called the *first and second supplements* to the law of quadratic reciprocity.

It should be noted that, if the congruence $x^2 \equiv a \bmod n$ is soluble then $(a/n) = 1$, but the converse need not hold when n is not prime. For example, if $n = 21$ and $a = 5$ then the congruence $x^2 \equiv 5 \bmod 21$ is insoluble, since both the congruences $x^2 \equiv 5 \bmod 3$ and $x^2 \equiv 5 \bmod 7$ are insoluble, but

$$\left(\frac{5}{21}\right) = \left(\frac{5}{3}\right)\left(\frac{5}{7}\right) = (-1)^2 = 1.$$

The Jacobi symbol finds an interesting application in the proof of the following result:

Proposition 8 *If a is an integer which is not a perfect square, then there exist infinitely many primes p not dividing a for which $(a/p) = -1$.*

Proof Suppose first that $a = -1$. Since $(-1/p) = (-1)^{(p-1)/2}$, we wish to show that there are infinitely many primes $p \equiv 3 \bmod 4$. Clearly 7 is such a prime. Let $\{p_1, \ldots, p_m\}$ be any finite set of such primes greater than 3. Adapting Euclid's proof of the infinity of primes (which is reproduced at the beginning of Chapter IX), we put

$$b = 4p_1 \cdots p_m + 3.$$

Then b is odd, but not divisible by 3 or by any of the primes $p_1, \ldots, p_m$. Since $b \equiv 3 \bmod 4$, at least one prime divisor q of b must satisfy $q \equiv 3 \bmod 4$. Thus the set $\{3, p_1, \ldots, p_m\}$ does not contain all primes $p \equiv 3 \bmod 4$.

Suppose next that $a = \pm 2$. Then $(a/5) = -1$. Let $\{p_1, \ldots, p_m\}$ be any finite set of primes greater than 3 such that $(a/p_i) = -1$ $(i = 1, \ldots, m)$ and put

$$b = 8p_1 \cdots p_m \pm 3,$$

where the $\pm$ sign is chosen according as $a = \pm 2$. Then b is not divisible by 3 or by any of the primes $p_1, \ldots, p_m$. Since $b \equiv \pm 3 \bmod 8$, we have $(2/b) = -1$ and $(a/b) = -1$ in both cases. If $b = q_1 \cdots q_n$ is the representation of b as a product of primes (repetitions allowed), then

$$(a/b) = (a/q_1) \cdots (a/q_n)$$

and hence $(a/q_j) = -1$ for at least one j. Consequently the result holds also in this case.

Consider now the general case. We may assume that a is square-free, since if $a = a'b^2$, where a' is square-free, then $(a/p) = (a'/p)$ for every prime p not dividing a. Thus we can write

$$a = \varepsilon 2^e r_1 \cdots r_h,$$

where $\varepsilon = \pm 1$, $e = 0$ or 1, and $r_1, \ldots, r_h$ are distinct odd primes. By what we have already proved, we may assume $h \geq 1$.

Let $\{p_1, \ldots, p_m\}$ be any finite set of odd primes not containing any of the primes $r_1, \ldots, r_h$. By Proposition 6, there exists an integer c such that $(c/r_1) = -1$. Since the moduli are relatively prime in pairs, by Corollary II.38 the simultaneous congruences

$$\begin{aligned} x &\equiv 1 \bmod 8, \quad & x &\equiv 1 \bmod p_i (i = 1, \ldots, m), \\ x &\equiv c \bmod r_1, \quad & x &\equiv 1 \bmod r_j (j = 2, \ldots, h), \end{aligned}$$

have a positive solution $x = b$. Then b is not divisible by any of the odd primes $p_1, \ldots, p_m$ or $r_1, \ldots, r_h$. Moreover $(-1/b) = (2/b) = 1$, since $b \equiv 1 \bmod 8$. Since $(r_j/b) = (b/r_j)$ for $1 \leq j \leq h$, it follows that

$$\begin{aligned} (a/b) &= (\varepsilon/b)(2/b)^e (r_1/b) \cdots (r_h/b) \\ &= (b/r_1)(b/r_2) \cdots (b/r_h) = (c/r_1)(1/r_2) \cdots (1/r_h) = -1. \end{aligned}$$

As in the special case previously considered, this implies that $(a/q) = -1$ for some prime q dividing b, and the result follows. □

A second proof of the law of quadratic reciprocity will now be given. Let p be an odd prime and, for any integer a not divisible by p, with Legendre *define*

$$(a/p) = 1 \quad \text{or} \quad -1$$

according as a is a quadratic residue or quadratic nonresidue of p. It follows from Euler's criterion (Proposition II.28) that

$$(ab/p) = (a/p)(b/p)$$

for any integers a, b not divisible by p. Also, by Corollary II.29,

$$(-1/p) = (-1)^{(p-1)/2}.$$

Now let q be an odd prime distinct from p and let $K = \mathbb{F}_q$ be the finite field containing q elements. Since $p \neq q$, the polynomial $t^p - 1$ has no repeated factors in K and thus has p distinct roots in some field $L \supseteq K$. If ζ is any root other than 1, then the (cyclotomic) polynomial

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

has the roots $\zeta^k (k = 1, \ldots, p-1)$.

Consider the *Gauss sum*

$$\tau = \sum_{x=1}^{p-1} (x/p)\zeta^x.$$

Instead of summing from 1 to $p-1$, we can just as well sum over any set of representatives of $\mathbb{F}_p^\times$:

$$\tau = \sum_{x \not\equiv 0 \bmod p} (x/p)\zeta^x.$$

Since q is odd, $(x/p)^q = (x/p)$ and hence, since L has characteristic q,

$$\tau^q = \sum_{x \not\equiv 0 \bmod p} (x/p)\zeta^{xq}.$$

If we put $y = xq$ then, since

$$(x/p) = (q^2 x/p) = (qy/p) = (q/p)(y/p),$$

we obtain

$$\tau^q = \sum_{y \not\equiv 0 \bmod p} (q/p)(y/p)\zeta^y = (q/p)\tau.$$

Furthermore,

$$\tau^2 = \sum_{u,v \not\equiv 0 \bmod p} (u/p)(v/p)\zeta^u \zeta^v = \sum_{u,v \not\equiv 0 \bmod p} (uv/p)\zeta^{u+v}$$

or, putting $v = uw$,

$$\tau^2 = \sum_{w \not\equiv 0 \bmod p} (w/p) \sum_{u \not\equiv 0 \bmod p} \zeta^{u(1+w)}.$$

Since the coefficients of t^{p-1} and t^{p-2} in $f(t)$ are 1, the sum of the roots is -1 and thus

$$\sum_{u \not\equiv 0 \bmod p} \zeta^{au} = -1 \quad \text{if } a \not\equiv 0 \bmod p.$$

On the other hand, if $a \equiv 0 \bmod p$, then $\zeta^{au} = 1$ and

$$\sum_{u \not\equiv 0 \bmod p} \zeta^{au} = p - 1.$$

Hence

$$\tau^2 = (-1/p)(p-1) - \sum_{w \not\equiv 0, -1 \bmod p} (w/p) = (-1/p)p - \sum_{w \not\equiv 0 \bmod p} (w/p).$$

Since there are equally many quadratic residues and quadratic nonresidues, the last sum vanishes and we obtain finally

$$\tau^2 = (-1)^{(p-1)/2} p.$$

Thus $\tau \neq 0$ and from the previous expression for τ^q we now obtain

$$\tau^{q-1} = (q/p).$$

But

$$\tau^{q-1} = (\tau^2)^{(q-1)/2} = \{(-1)^{(p-1)/2} p\}^{(q-1)/2}$$

and $p^{(q-1)/2} = (p/q)$, by Proposition II.28 again. Hence

$$(q/p) = (-1)^{(p-1)(q-1)/4} (p/q),$$

which is the law of quadratic reciprocity.

The preceding proof is a variant of the sixth proof of Gauss (1818). Already in 1801 Gauss had shown that if p is an odd prime, then

$$\sum_{k=0}^{p-1} e^{2\pi i k^2/p} = \pm\sqrt{p} \text{ or } \pm i\sqrt{p} \text{ according as } p \equiv 1 \text{ or } p \equiv 3 \bmod 4.$$

After four more years of labour he managed to show that in fact the $+$ signs must be taken. From this result he obtained his fourth proof of the law of quadratic reciprocity. The sixth proof avoided this sign determination, but Gauss's result is of interest in itself. Dirichlet (1835) derived it by a powerful analytic method, which is readily generalized. Although we will make no later use of it, we now present Dirichlet's argument.

For any positive integers m, n, we define the *Gauss sum* $G(m, n)$ by

$$G(m,n) = \sum_{v=0}^{n-1} e^{2\pi i v^2 m/n}.$$

Instead of summing from 0 to $n-1$ we can just as well sum over any complete set of representatives of the integers mod n:

$$G(m,n) = \sum_{v \bmod n} e^{2\pi i v^2 m/n}.$$

Gauss sums have a useful multiplicative property:

Proposition 9 *If m, n, n' are positive integers, with n and n' relatively prime, then*

$$G(mn', n)G(mn, n') = G(m, nn').$$

Proof When v and v' run through complete sets of representatives of the integers mod n and mod n' respectively, $\mu = vn' + v'n$ runs through a complete set of representatives of the integers mod nn'. Moreover

$$\mu^2 m = (vn' + v'n)^2 m \equiv (v^2 n'^2 + v'^2 n^2)m \bmod nn'.$$

It follows that

$$\begin{aligned} G(mn', n)G(mn, n') &= \sum_{v \bmod n} \sum_{v' \bmod n'} e^{2\pi i (mn'^2 v^2 + mn^2 v'^2)/nn'} \\ &= \sum_{\mu \bmod nn'} e^{2\pi i \mu^2 m/nn'} = G(m, nn'). \end{aligned}$$

□

A deeper result is the following reciprocity formula, due to Schaar (1848):

Proposition 10 *For any positive integers m, n,*

$$G(m,n) = \sqrt{\frac{n}{m}} C \sum_{\mu=0}^{2m-1} e^{-\pi i \mu^2 n/2m},$$

where $C = (1+i)/2$.

Proof Let $f : \mathbb{R} \to \mathbb{C}$ be a function which is continuously differentiable when restricted to the interval $[0, n]$ and which vanishes outside this interval. Since the sum

$$F(t) = \sum_{k=-\infty}^{\infty} f(t+k)$$

has only finitely many nonzero terms, the function F has period 1 and is continuously differentiable, except possibly for jump discontinuities when t is an integer. Therefore,

by Dirichlet's convergence criterion in the theory of Fourier series,

$$\{F(+0)+F(-0)\}/2=\lim_{N\to\infty}\sum_{h=-N}^{N}\int_0^1 e^{-2\pi iht}F(t)dt.$$

But

$$\begin{aligned}\int_0^1 e^{-2\pi iht}F(t)dt&=\sum_{k=-\infty}^{\infty}\int_0^1 e^{-2\pi iht}f(t+k)dt\\&=\sum_{k=-\infty}^{\infty}\int_k^{k+1} e^{-2\pi iht}f(t)dt=\int_0^n e^{-2\pi iht}f(t)dt.\end{aligned}$$

Thus we obtain

$$f(0)/2+f(1)+\cdots+f(n-1)+f(n)/2=\lim_{N\to\infty}\sum_{h=-N}^{N}\int_0^n e^{-2\pi iht}f(t)dt. \quad (*)$$

This is a simple form of *Poisson's summation formula* (which makes an appearance also in Chapters IX and X).

In particular, if we take $f(t)=e^{2\pi it^2m/n}$ $(0\le t\le n)$, where m is also a positive integer, then the left side of $(*)$ is just the Gauss sum $G(m,n)$. We will now evaluate the right side of $(*)$ for this case. Put $h=2mq+\mu$, where q and μ are integers and $0\le\mu<2m$. Then

$$e^{-2\pi iht}f(t)=e^{2\pi im(t-nq)^2/n}e^{-2\pi i\mu t}.$$

As h runs through all the integers, q does also and μ runs independently through the integers $0,\dots,2m-1$. Hence

$$\begin{aligned}\lim_{N\to\infty}\sum_{h=-N}^{N}\int_0^n e^{-2\pi iht}f(t)dt&=\sum_{\mu=0}^{2m-1}\lim_{Q\to\infty}\sum_{q=-Q}^{Q}\int_0^n e^{2\pi im(t-nq)^2/n}e^{-2\pi i\mu t}dt\\&=\sum_{\mu=0}^{2m-1}\lim_{Q\to\infty}\sum_{q=-Q}^{Q}\int_{-qn}^{-(q-1)n} e^{2\pi it^2m/n}e^{-2\pi i\mu t}dt\\&=\sum_{\mu=0}^{2m-1}\int_{-\infty}^{\infty} e^{2\pi it^2m/n}e^{-2\pi i\mu t}dt\\&=\sum_{\mu=0}^{2m-1}\int_{-\infty}^{\infty} e^{2\pi im(t-\mu n/2m)^2/n}e^{-\pi i\mu^2n/2m}dt\\&=\sum_{\mu=0}^{2m-1} e^{-\pi i\mu^2n/2m}\int_{-\infty}^{\infty} e^{2\pi it^2m/n}dt\\&=\sqrt{\frac{n}{m}}C\sum_{\mu=0}^{2m-1} e^{-\pi i\mu^2n/2m},\end{aligned}$$